Book of Scientific Abstracts

SUN OCT 1

7:00 AM – 7:45 AM ET	Registration & Continental Breakfast
7:45 AM – 8:00 AM ET	Welcome Remarks
8:00 AM – 9:00 AM ET	Opening Keynote Address Expanding Horizons: An In-depth Exploration of Generative AI in Medical Imaging
9:00 AM – 10:30 AM ET	**Scientific Abstract Presentations: Clinical Applications**
10:30 AM – 10:45 AM ET	Break
10:45 AM – 12:15 PM ET	**Scientific Abstract Presentations: Intelligent Imaging**
12:15 – 1:15 PM ET	Lunch & Visit Scientific Abstract Posters
1:15 PM – 2:45 PM ET	Town Hall: Synthesizing Diagnostic Imaging Data Scientists
2:45 PM – 4:15 PM ET	**Scientific Abstract Presentations: Data Sets, Emerging Technologies & NLP Models**
4:15 PM – 4:30 PM ET	Break
4:30 PM – 5:15 PM ET	The Do's and Don'ts of Publishing Machine Learning Manuscripts in the *Journal of Digital Imaging*
5:15 PM – 6:30 PM ET	Choosing the Right Platform for Your AI Applications - A Vendor Panel Discussion
6:30 PM – 7:45 PM ET	**Networking Reception & Scientific Abstract Poster Discussion**

MON OCT 2

7:00 AM – 7:45 AM ET	Registration & Continental Breakfast
7:45 AM – 8:45 AM ET	AAPM-SIIM Symposium on Machine Intelligence in Medical Imaging: A Medical Physics Perspective
8:45 AM – 10:15 AM ET	**Scientific Abstract Presentations: Generative AI & General Applications in NLP**
10:15 AM – 10:30 AM ET	Break
10:30 AM – 12:00 PM ET	**Scientific Abstract Presentations: Toolkits and Machine Learning Algorithms**
12:00 PM – 1:00 PM ET	Lunch & Visit Scientific Abstract Posters
1:00 PM – 2:15 PM ET	**Scientific Abstract Presentations: Clinical Applications**
2:15 PM – 3:15 PM ET	Regulatory Science for AI-enabled Devices in Medical Imaging: A Perspective from the AI/ML Regulatory Science Research Program at the Office of Science and Engineering Laboratories (OSEL) at the FDA
3:15 PM ET	Closing Remarks

Scientific Abstract Presentations: Clinical Applications

Date:
SUN, OCT 1

Time:
9:00 AM – 10:30 AM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

A Comparison of Convolutional Neural Network Architectures in Auto-Segmenting Primary Oropharyngeal Cancers from Contrast-Enhanced CT Scans
+ Onur Sahin, PhD, Medical Student, McGovern Medical School
+ Kareem A. Wahid, PhD; Abdallah S. Mohamed, MD, PhD; Clifton D. Fuller, MD, PhD; Mohamed A. Naser, PhD

Acute Respiratory Distress Syndrome (ARDS) Detection in the Pediatric Intensive Care Unit (PICU) setting Demonstrates High Performance with Transfer Deep Learning
+ Vahid Khalkhali, PhD, Research Scientist, Children's Hospital of Philadelphia
+ Michael Welsh, DO; Patricia P. Rafful, MD, PhD; Dana S. Alkhulaifat, MD; Adarsh Ghosh, MD; Saurav Bose, MD; Nadir Yehya, MD; Susan T. Sotardi, MD

An Improved UNet++ Architecture for Deep Learning based Segmentation of Kidneys and Cysts in Autosomal Dominant Polycystic Kidney Disease (ADPKD)
+ Chetana Krishnan, Graduate Student, University of Alabama at Birmingham
+ Emma Schmidt; Ezinwanne Onuoha, MS; Michal Mrug, MD; Carlos E Cardenas, PhD; Harrison Kim, PhD

Augmenting the MIDRC Dataset using Deep Learning-Based Quantification of Abdominal Aortic Calcification: Proof-of-Concept for Population-Level Disease Screening
+ Devina Chatterjee, Medical Student, University of Maryland School of Medicine;
+ Adway Kanhere, MS; Annie Trang, MS; Vishwa S. Parekh, PhD; Paul H. Yi, MD, MS

Deep Multiclass Multiple-instance Learning For DSA Classification
+ Reza Moein Taghavi, Medical Student, UC Davis School of Medicine
+ Roger Goldman, MD, PhD

Evaluation of an Artificial Intelligence Chatbot for Delivery of Interventional Radiology Patient Education Material
+ Colin J. McCarthy, MD, Interventional Radiologist, Beth Israel Deaconess Medical Center, Harvard Medical School
+ Seth Berkowitz, MD; Vijay Ramalingam, MD; Muneeb Ahmed, MD

A Comparison of Convolutional Neural Network Architectures in Auto-Segmenting Primary Oropharyngeal Cancers from Contrast-Enhanced CT Scans

Onur Sahin, PhD, Medical Student, McGovern Medical School; Kareem A. Wahid, PhD; Abdallah S. Mohamed, MD, PhD; Clifton D. Fuller, MD, PhD; Mohamed A. Naser, PhD

Introduction

Contouring tumors during the development of radiotherapy plans is a labor-intensive process with a high degree of variation when manually performed. Deep learning (DL) models have been developed to automate the segmentation of tumors from anatomical imaging modalities. Previously, our group developed an auto-segmentation model for oropharyngeal cancers (OPCs) from MRI scans. While multiple studies focused on auto-segmentation models from other imaging modalities, there has been no work assessing the development of models using diagnostic-quality contrast-enhanced computed tomography (CE-CT) scans for OPCs , even though CE-CTs can be more easily acquired. In this study, we develop two models using Swin and Resunet convolutional neural networks to auto-segment OPCs from CE-CT scans.

Hypothesis

We hypothesize that DL convolution neural networks can be trained on CE-CT scans to segment OPCs.

Methods

Pre-surgical diagnostic quality CE-CT scans were collected from 474 patients with OPCs. Ground-truth segmentations of OPCs were manually performed and images were randomly assigned into a training (n=380) and test set (n=94) using a 4:1 split. 5-fold cross validation was performed on the training set to train 5 separate models using either a Resunet or Swin network from the MONAI framework. Final model predictions were generated by a consensus mask using either an averaging or simultaneous truth and performance level estimation (STAPLE) method and assessed on the test set. The Dice similarity coefficient (DSC) was used to evaluate model performance. Statistical comparisons were performed using Wilcoxon signed-ranks test.

Results

In the training/test sets, 16.8%/17% of OPCs were T1, 42.1%/42.5% were T2, 22.3%/22.3% were T3, and 18.6%/18.1% were T4. The overall average DSC performance across cross-validation folds was 0.61/0.62 for the Swin/Resunet models, 0.71/0.7 consensus masks developing via an averaging method on the test set, and 0.7/0.69 for the STAPLE method. The Wilcoxon signed-ranks test between the Swin and Resunet showed a $p < 0.05$ for only the consensus masking generated via the averaging method.

Conclusion

We developed the first DL models, to our knowledge, trained on CE-CT scans to segment OPCs, and showed that

they achieved similar performance to DL models trained on MRI images. Further work can focus on improving segmentation performance by utilizing larger patient cohort sizes.

Statement of Impact

In this study we show that more CE-CT scans can be used to develop OPC auto-segmentation models with reasonable performance. With further optimization, the proposed models could be integrated into the workflow of radiation oncologists to increase the efficiency of developing radiation treatment plans.

Keywords

Auto-segmentation; Oropharyngeal Cancer; Convolutional Neural Networks

Acute Respiratory Distress Syndrome (ARDS) Detection in the Pediatric Intensive Care Unit (PICU) setting Demonstrates High Performance with Transfer Deep Learning

Vahid Khalkhali, PhD, Research Scientist, Children's Hospital of Philadelphia; Michael Welsh, DO; Patricia P. Rafful, MD, PhD; Dana S. Alkhulaifat, MD; Adarsh Ghosh, MD; Saurav Bose, MD; Nadir Yehya, MD; Susan T. Sotardi, MD

Introduction

Acute respiratory distress syndrome (ARDS) is a significant cause of morbidity and mortality in the pediatric intensive care unit (PICU). ARDS diagnosis involves chest X-ray (CXR) criteria combined with clinical and laboratory parameters. Machine learning models have demonstrated utility in the detection of ARDS on chest radiographs.

Hypothesis

Machine learning models for radiology CXR can perform better than humans. We evaluated the performance of deep learning (DL) models to diagnose ARDS based on CXR exams. We performed statistical evaluation performance between the models and two experienced radiologists, and also statistical agreement evaluation between the two radiologists as a gold-standard correlation for comparison.

Methods

In this retrospective, IRB-approved study, we identified 368 children admitted to the PICU with a diagnosis of ARDS, at a large pediatric academic center from 2014 to 2019. A single random radiograph from all patients admitted to the PICU without a diagnosis of ARDS during 2018 was used as the control cohort, (n=1127). The train-validation-test ratio was 60/20/20. Using transfer learning, we utilized pretrained DL structures to diagnose ARDS (PyTorch, version 1.2). The area under the receiver operating characteristic (AUROC) was the main performance metric. Two pediatric radiologists independently assigned labels of ARDS/No ARDS and interrater reliability was calculated. Correlations were calculated using the Pearson correlation coefficient and Cohen’s Kappa. All statistical analyses used Type-I error of 5% and power of 80%.

Results

The interrater reliability between the radiologists was 94.5% (Cohen’s Kappa of 85.8%) for the training cohort. The ARDS diagnostic performance of two radiologists yielded an AUROC (balanced accuracy) of 72.5%, while the DenseNet161 model achieved 86.0% (AUROC of 92.5%) and an ensemble of models reach 83.7% (AUROC of 93.5%). Radiologist diagnoses were only 81% correlated (Cohen’s Kappa 51.6%) with the DenseNet161 and 83% (Cohen’s Kappa 60.5%) with the ensemble. While the Pearson correlation between the two radiologists was high (>90%) on the test set, the difference between the detection of different models was statistically significant ($p < 0.01$). Attention maps show that models are able to capture the regions of interest (Figure 1).

Conclusion

Using transfer DL, we trained models to reliably detect ARDS in the PICU and compared their performance with the diagnostic rates of two experienced radiologists. DL can automatically detect ARDS on chest radiographs, with a performance that parallels those of radiologists.

Statement of Impact

Deep learning detection of ARDS could improve the triage of patients in the intensive care unit before the availability of dedicated pediatric radiologist reads.

Figure 1. Attention maps for CNN: left and right examples are non-ARDS and ARDS, respectively.

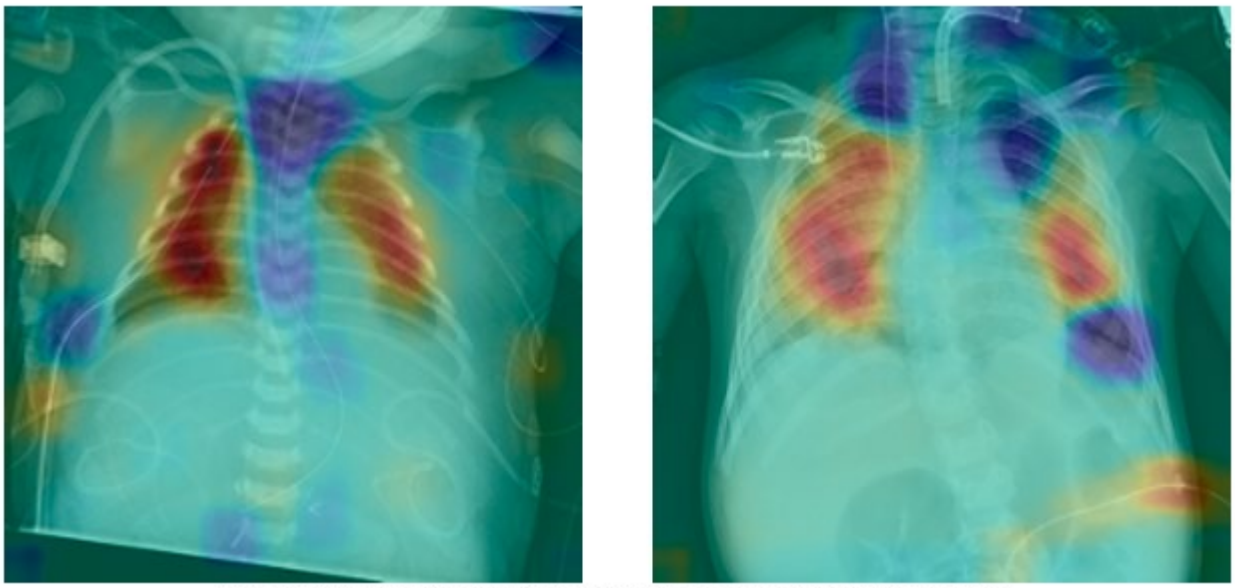


Figure 1. Attention maps for CNN: left and right examples are non-ARDS and ARDS, respectively.

Keywords

Acute Respiratory Distress Syndrome (ARDS)**;** Transfer Deep Learning**;** Chest X-Ray (CXR)

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

An Improved UNet++ Architecture for Deep Learning based Segmentation of Kidneys and Cysts in Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Chetana Krishnan, Graduate Student, University of Alabama at Birmingham; Emma Schmidt; Ezinwanne Onuoha, MS; Michal Mrug, MD; Carlos E Cardenas, PhD; Harrison Kim, PhD

Introduction

UNet++, an enhanced version of UNet, incorporates nested skip connections (NSC), batch normalization (BN), and deep supervision to improve feature extraction and accuracy. However, the UNet++ encoder lacks sufficient feature extraction in small regions of interest (ROIs), leading to inadequate feature fusion during up-sampling and reduced accuracy due to the absence of proper feature normalization. We propose a new architecture called sUNet++, replacing BN with switching normalization (SN) to avoid batch effects and integrate residual blocks in the encoder and decoder. NSCs are replaced with concatenated skip connections (CSC). We compared this architecture with gUNet++, employing group normalization.

Hypothesis

Our hypothesis posits that SN identifies suitable normalization techniques for each layer based on their importance, while residual blocks propagate essential features through skip connections. CSCs improve gradient flow and information propagation. This mitigates performance degradation, allowing the network to learn residual mappings and retain information across layers, resulting in improved segmentation.

Methods

To validate our hypothesis, we trained UNet++, sUNet++, and gUNet++ on T2-weighted MRI images of 95 ADPKD patients, utilizing a total of 756 3D kidney images (604 for training, 76 for validation, and 76 for testing). Preprocessing, cropping, and slicing techniques were applied to generate 2D training samples, resulting in approximately 69,000 samples. The task involved segmenting kidneys and cysts. The models were trained for 50 epochs using a patch-wise approach. Data augmentation techniques were employed to increase the training samples. Leaky ReLU was the activation function. Performance was evaluated using the Dice similarity coefficient (DSC), Hausdorff distance (HD), and Intersection over Union score (IoU).

Results

As summarized in Table 1, sUNet++ achieved higher accuracy than UNet++ for both kidney and cyst segmentation. Moreover, sUNet++ exhibited the highest minimum dice score, indicating superior individual dice performance and success in cases where UNet++ failed. Although gUNet++ showed better performance for kidneys, it was less suitable for cyst segmentation. Notably, sUNet++ required fewer model parameters, converged faster, and demanded less training and inference time. Figure 1 shows the representative test kidney and cyst boundaries determined by sUNet++, gUNet++, and UNet++. Figure 2 illustrates the importance layers of sUNet++ and UNet++.

Conclusion

sUNet++ addresses the challenge of learning normalization in deep learning by dynamically selecting normalizations and statistics for each layer, offering architectural flexibility, and adaptability to varying batch sizes, and eliminating the reliance on sensitive hyperparameters.

Statement of Impact

sUNet++ can improve the diagnosis, treatment planning, and monitoring of ADPKD by detecting early changes in cysts.

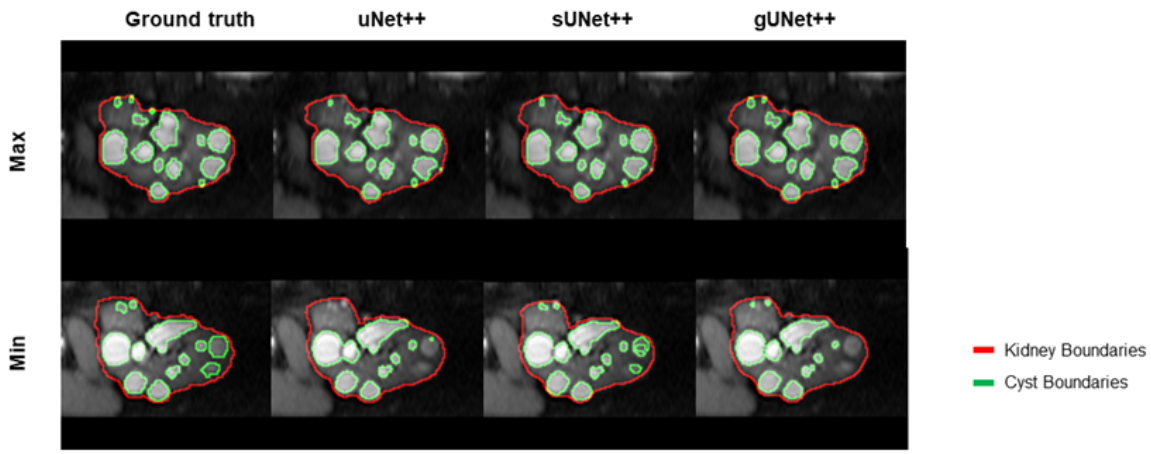


Figure 1. Ground truth Vs predicted kidney and cyst segmentation. Two representative images showing best (max) and moderate (min) performance on the test set, respectively, with kidney (red line) and cyst (green line) boundaries determined by our semi-automatic method (ground truth, first column) and all three models (second to fourth columns) sUNet++ demonstrates improved boundary, edge localization, and enhanced accuracy in capturing collided cyst edges. The replacement of concatenated skip connections enables the preservation of high-resolution edge information by establishing direct paths for information flow between corresponding encoder and decoder layers. This is advantageous for transferring learning to a different task.

Figure 1. Ground truth Vs predicted kidney and cyst segmentation. Two representative images showing best (max) and moderate (min) performance on the test set, respectively, with kidney (red line) and cyst (green line) boundaries determined by our semi-automatic method (ground truth, first column) and all three models (second to fourth columns) sUNet++ demonstrates improved boundary, edge localization, and enhanced accuracy in capturing collided cyst edges. The replacement of concatenated skip connections enables the preservation of high-resolution edge information by establishing direct paths for information flow between corresponding encoder and decoder layers. This is advantageous for transferring learning to a different task.

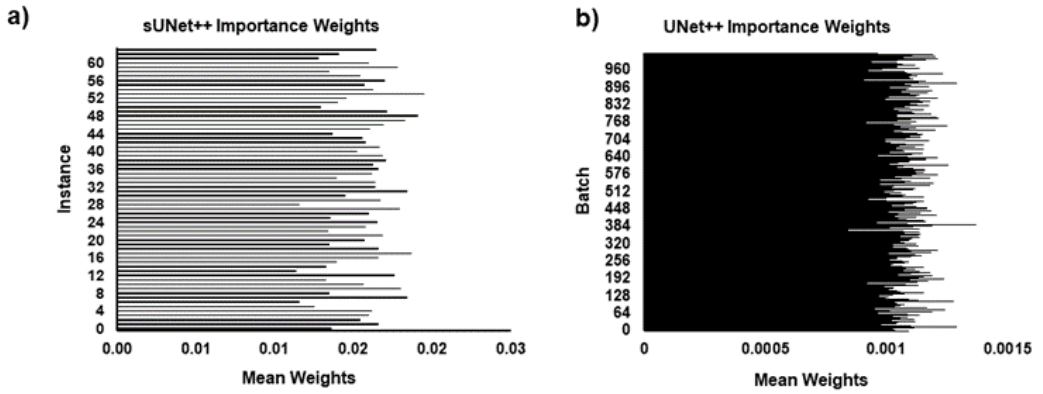


Figure 2. Importance weights. a) The model weight distribution per instance for sUNet++. b) The model weight distribution per batch for UNet++. sUNet++ has fewer model parameters and the concatenated skip connections combined all the layers with equal importance to one single layer to avoid complexity. This reduces redundancy and helps in faster convergence and training. UNet++ with batch normalization considers all the batches irrespective of importance thus leading to re-learning the same features causing overfitting on training. The mean weights are higher in sUNet++ for fewer parameters compared to UNet++ thus having improved representative learning thus removing batch effects.

Figure 2. Importance weights. a) The model weight distribution per instance for sUNet++. b) The model weight distribution per batch for UNet++. sUNet++ has fewer model parameters and the concatenated skip connections combined all the layers with equal importance to one single layer to avoid complexity. This reduces redundancy and helps in faster convergence and training. UNet++ with batch normalization considers all the batches irrespective of importance thus leading to re-learning the same features causing overfitting on training. The mean weights are higher in sUNet++ for fewer parameters compared to UNet++ thus having improved representative learning thus removing batch effects.

Model	ROI	IoU Score	HD (mm)	Test DSC	Min Test DSC	Max Test DSC	Training Time / Epoch (s)	Inference Time (min ± sec)
UNet++	Kidney	0.88±0.47	1.35±0.95	0.93±0.35	0.70	0.97	217±5	05±23
	Cyst	0.77±0.43	1.52±0.78	0.86±0.42	0.71	0.92	220±7	07±14
sUNet++	Kidney	0.90±0.44	1.38±0.94	0.94±0.35	0.84	0.98	110±3	04±12
	Cyst	0.77±0.47	1.30±0.90	0.87±0.42	0.76	0.93	116±6	04±34
gUNet++	Kidney	0.88±0.46	1.43±0.91	0.93±0.34	0.74	0.97	203±4	05±09
	Cyst	0.76±0.44	1.53±0.79	0.86±0.43	0.65	0.93	218±9	05±16

Table 1. Performance Metrics. Intersection over union (IoU) score, Haudsdorff distance (HD), and test Dice similarity score (DSC) of UNet++ and proposed models based kidney and cyst segmentation, together with its training time per epoch, and inference time to predict the boundaries of all test images.

Table 1. Performance Metrics. Intersection over union (IoU) score, Haudsdorff distance (HD), and test Dice similarity score (DSC) of UNet++ and proposed models based kidney and cyst segmentation, together with its training time per epoch, and inference time to predict the boundaries of all test images.

Keywords
Skip connections; Normalization; Image segmentation; Polycystic kidney disease; UNet++; Residual staging

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Augmenting the MIDRC Dataset using Deep Learning-Based Quantification of Abdominal Aortic Calcification: Proof-of-Concept for Population-Level Disease Screening

Devina Chatterjee, Medical Student, University of Maryland School of Medicine; Adway Kanhere, MS; Annie Trang, MS; Vishwa S. Parekh, PhD; Paul H. Yi, MD, MS

Introduction

Large public imaging datasets like the MIDRC dataset of 20,000+ CT scans have facilitated rapid development of machine learning tools to fight diseases like COVID-19. However, these datasets seldom have disease labels beyond their primary use case (e.g., COVID-19 status for MIDRC), limiting their use for other prediction tasks. We evaluated the feasibility of deep learning-based quantification of abdominal aortic calcification to augment the MIDRC dataset with potential biomarkers for population-level cardiovascular risk assessment.

Hypothesis

Deep learning-based abdominal aortic calcium quantification will allow for augmentation of the MIDRC dataset with population-level assessments of cardiovascular disease risk.

Methods

We first validated the state-of-the-art TotalSegmentator deep learning multi-organ CT segmentation model on two datasets of CT scans of the abdomen/pelvis for segmentation of the abdominal aorta and imaged portions of the thoracic aorta using Dice scores: 1) subset of the MIDRC dataset (COVID-19-NY-SBU [N=1285]; primary dataset) and 2) AMOS dataset (N=250; secondary dataset). Aortic calcifications were segmented on the COVID-19-NY-SBU dataset using Hounsfield Unit voxel thresholding (>250 HU); automated calcium segmentations were validated using 100 manually segmented scans. The aortic calcifications were localized to thoracic vs. abdominal aorta using an automated bodypart regressor model. Agatston score for the abdominal aorta was calculated by multiplying area of calcium by a factor related to maximum plaque attenuation.

Results

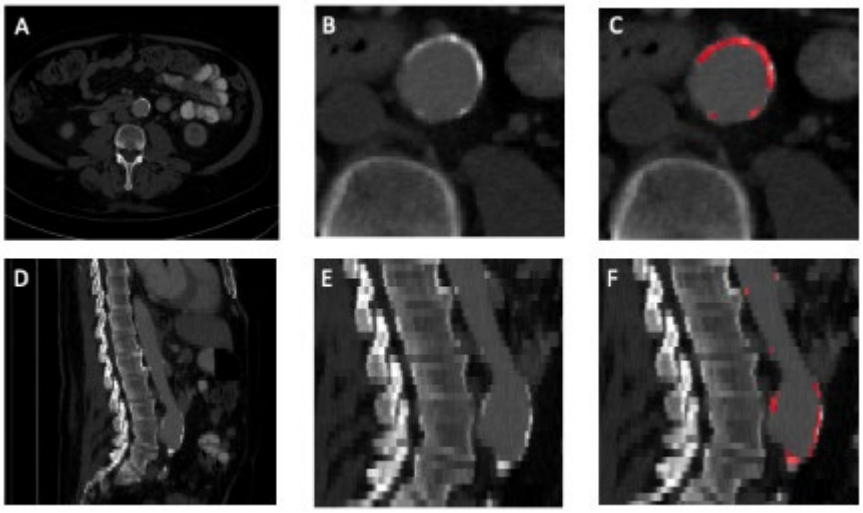
TotalSegmentator had high performance for aortic segmentation on both datasets with mean Dice scores of 0.88 and 0.91 on the MIDRC and AMOS datasets, respectively (Figure 1). Automated aortic calcification similarly had Dice score of 0.88 on our primary dataset (MIDRC) with high correlation with manual segmentations (R2 of 0.9575; Figure 2). In the MIDRC dataset, 75% of patients had aortic calcification; of these, 73% had abdominal aortic calcifications and 77% had thoracic aortic calcifications. For patients with calcifications [N=963], the mean Agatston score in the abdominal aorta was 15,235; 792 patients (62%) had scores >1000 (previously-validated cardiovascular disease risk threshold).

Conclusion

Deep learning-based quantification of abdominal aortic calcifications can augment large public datasets like MIDRC, allowing for population-level assessments of cardiovascular disease risk. Our proof-of-concept paves the way for augmenting other large datasets, including the remainder of the MIDRC dataset, which our group is actively working on next.

Statement of Impact

Augmentation of large imaging datasets for disease risk assessment with deep learning models paves the way for population-level screening for chronic conditions like cardiovascular disease.



Examples of deep learning-based aortic calcification segmentation in the abdominal aorta. Panels (A) and (D) show single axial and sagittal slices, respectively. Panels (C) and (G) demonstrate a portion of the aorta with calcium. Panels (D) and (H) highlight in red the aortic calcification automatically detected by the deep learning model. Note that our models segmented calcium in the entire 3D CT volume, from which the Agatston scores were calculated.

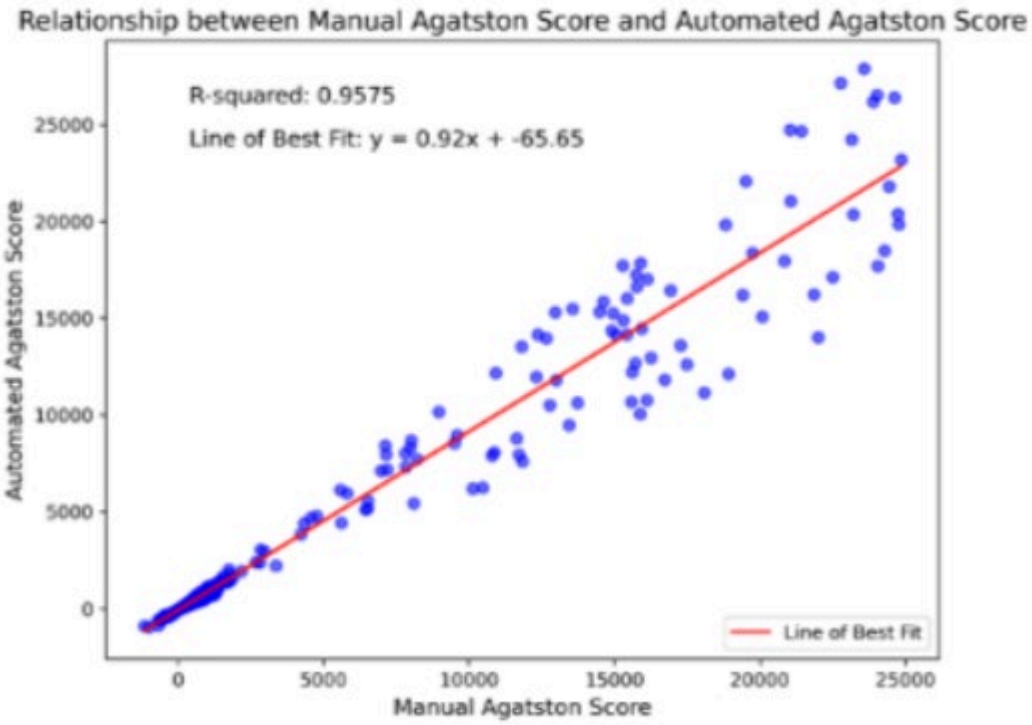


Figure 2: Deep learning-based aortic calcification segmentation and Agatston score calculation correlates well with manual segmentation and score calculations with correlation coefficient of 0.96.

Keywords

MIDRC; Datasets; Aortic Calcification; Opportunistic Screening; Deep Learning

Deep Multiclass Multiple-instance Learning For DSA Classification

Reza Moein Taghavi, Medical Student, UC Davis School of Medicine; Roger Goldman, MD, PhD

Introduction
Anatomic localization is a critical requirement for interpretation of angiography. Many images in digital subtraction angiography (DSA) sequences lack information for localization due to insufficient radiopaque contrast within the vessels, presenting a significant challenge in automated interpretation. The purpose of our study was to evaluate a deep multiclass Multiple Instance Learning (MIL) algorithm for anatomic localization in DSA sequences.

Hypothesis
We hypothesize that MIL model can accurately identify standard anatomic locations in abdominopelvic angiographic sequences.

Methods
We performed a retrospective review of the institutional PACS to identify 689 DSA sequences performed with contrast administration via the aorta, left external iliac artery, right external iliac artery, celiac artery, superior mesenteric artery, and inferior mesenteric artery. Individual images within each DSA sequence were designated as “key” if contrast opacified the identified artery at the location of contrast administration and a first order downstream vessel. Data were divided into 482 sequences for training and validation and 207 sequences for testing. A deep multiclass MIL model was developed using the MONAI Python library to classify DSA sequences from these anatomical locations. The model was trained with inputs of 50 images from each angiographic sequence. To ensure uniformity, all sequences were padded to 50 images with copies of the final image. Classification performance was quantified using accuracy, precision, recall, and F1. MIL model assigned attention weights to each image reflecting contribution to the final classification. Images corresponding to the algorithmically generated top five attention weights were compared with manually-labeled “key” images for overlap. The overlap was quantified as the ratio of the number of images in common to the number of key images.

Results
The deep multiclass MIL algorithm achieved an accuracy of 92.75% (95% CI: 89.22 - 96.28), Precision: 93.99% ± 3.24, Recall: 92.75% ± 3.53, and F1 of 88% ± 4.43 on the held-out test data. The algorithm performance for each anatomical location is provided in table 1. Figure 1 depicts an example of a DSA image sequence, manually-labeled “key” images, and the attention weights. We found an average overlap of 54.8%. In 93.24% of cases, at least one algorithm-chosen image matched an image in the manually-labeled “key” selections.

Conclusion
Deep multiclass MIL is feasible for accurate anatomic localization in DSA imaging.

Statement of Impact
This study demonstrates the potential of deep learning with attention mechanisms for automatically classifying the anatomical locations in time-series DSA data, a critical task in the interpretation of imaging during and after image-guided endovascular procedures.

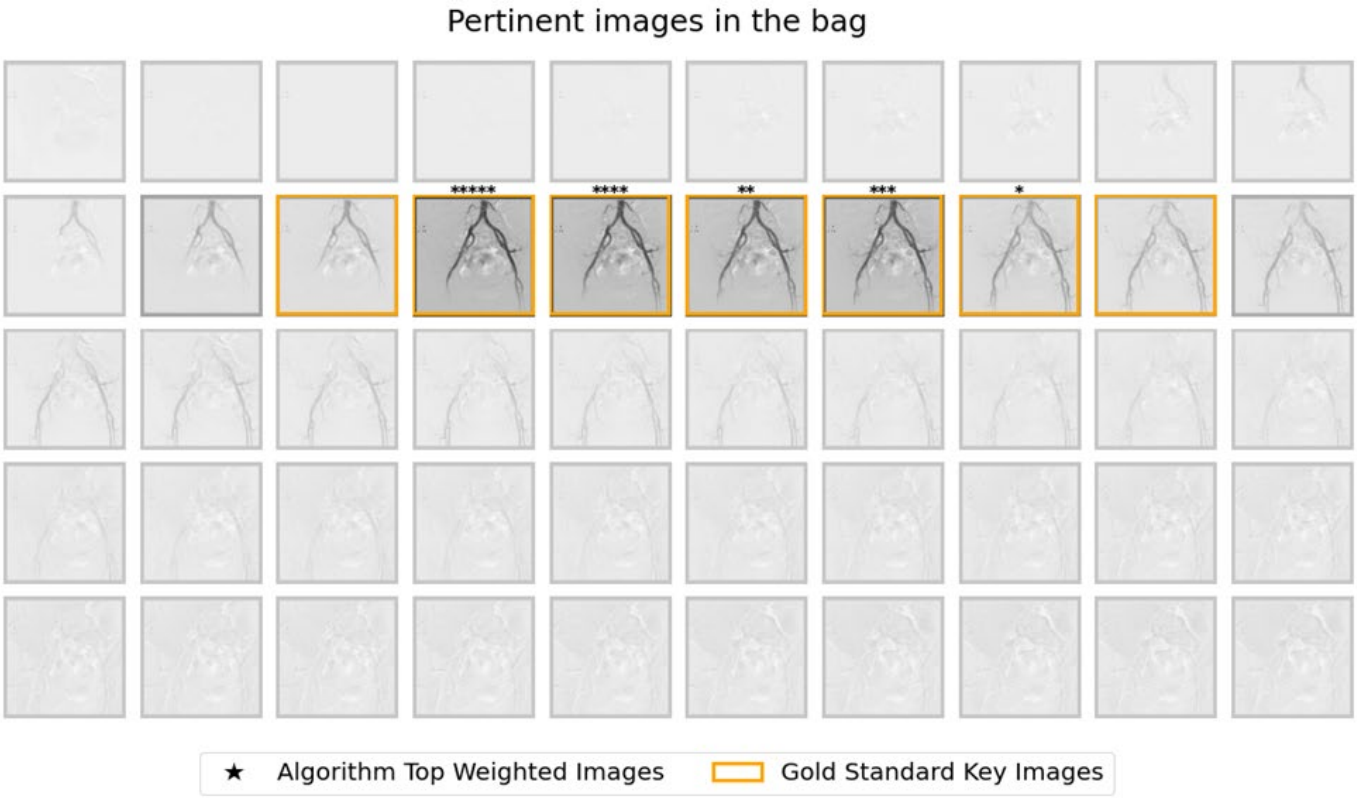


Figure 1: Sample DSA sequence from the testing dataset with an overlay of algorithm-selected images and manually labeled diagnostic “key” images are shown. Images with stars represent the top five weighted images by the algorithm, with the star count correlating to the assigned weight. Similarly, the opacity of the images correlates to algorithm generated weights: the greater the opacity, the larger the weight. Conversely, more transparent images have smaller weights. Images outlined in yellow rectangles represent the diagnostic “key” images.
HAPYGCNX-1580149-2-ANY(4).docx

Anatomic location	Number of samples	Precision	Sensitivity (Recall)	F1
Aorta	27	100	96.3	98.12
CA	66	95.59	98.48	97.01
EIA/R	68	88.16	98.53	93.06
EIA/L	13	100	69.23	81.82
SMA	28	88	78.57	83.02
IMA	5	100	60	75

Table 1 : Performance of the algorithm on the test dataset, stratified by anatomical location: aorta (AO), left external iliac artery (LEIA), right external iliac artery (REIA), celiac artery (CA), superior mesenteric artery (SMA), and inferior mesenteric (IMA).

Keywords
MIL; Deep Learning; DSA

Evaluation of an Artificial Intelligence Chatbot for Delivery of Interventional Radiology Patient Education Material

Colin J. McCarthy, MD, Interventional Radiologist, Beth Israel Deaconess Medical Center, Harvard Medical School; Seth Berkowitz, MD; Vijay Ramalingam, MD; Muneeb Ahmed, MD

Introduction

To assess the potential role for ChatGPT for the delivery of medical information to patients.

Hypothesis

To assess the accuracy, completeness, and readability of patient educational material produced by a machine-learning model and compare the output to that provided by a Societal patient education website.

Methods

Content from the Society of Interventional Radiology (SIR) Patient Center website was retrieved, categorized and organized into discrete questions. These questions were entered into the ChatGPT platform, and the output was analyzed for word and sentence count, readability using multiple validated scales, factual correctness and suitability for patient education using the PEMAT-P instrument.

Results

21,154 words were analyzed, including 7,917 words from the website and 13,377 words representing the total output of the ChatGPT platform across twenty-two text passages. Compared to the Societal website, output from the ChatGPT platform was longer and more difficult to read on 4 of 5 readability scales. The ChatGPT output was incorrect for 12 of 104 (11.5%) questions. When reviewed using the PEMAT-P tool, the ChatGPT content scored lower than the website material. Content from both the website and ChatGPT were significantly above the recommended 5th or 6th grade-level for patient education, with mean Flesch Kincaid Grade Level of 11.1 (+/- 1.3) for the website and 11.9 (+/- 1.6) for the ChatGPT content.

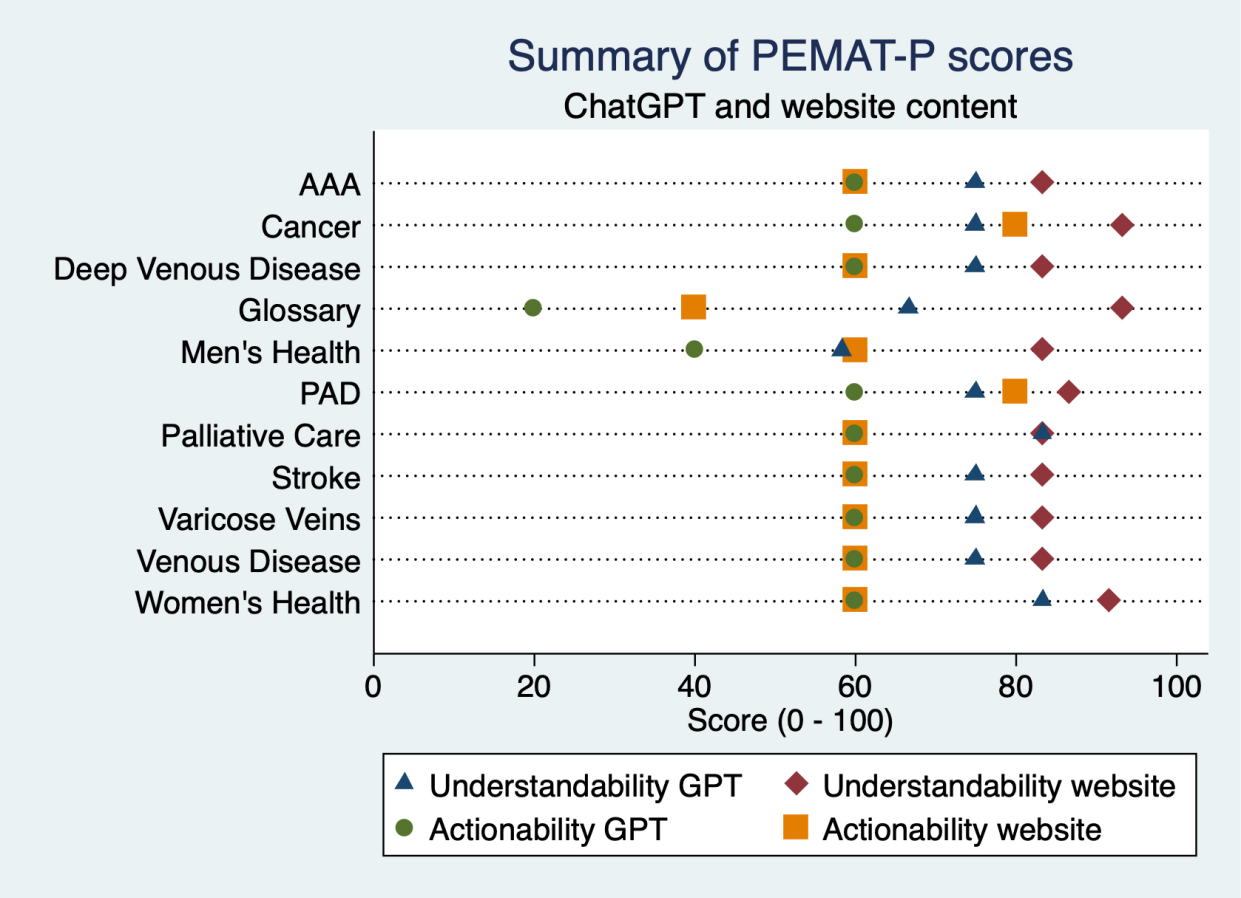
Conclusion

The ChatGPT platform may produce incomplete or inaccurate patient educational content, and providers should be familiar with the limitations of the system in its current form. Opportunities may exist to fine-tune existing large language models, which could be optimized for the delivery of patient educational content.

Statement of Impact

1. Evaluation of this large language Chatbot model as a patient education tool showed that incomplete or incorrect content was provided in 12 of 104 (11.5%) questions. 2. The absence of visual aids in the generated output results in material that is not fully optimized for patient educational purposes. 3. Compared to the Societal website, content from the ChatGPT platform was longer, contained more difficult words and longer sentences, and was more difficult

to read when assessed using several validated reading scales. Content from both the ChatGPT model and Societal website was written at a grade level higher than that recommended for patient education.



Performance of the ChatGPT model against the SIR website, using the PEMAT-P instrument.

Table 3 – overview of analyzed text passages.

	Word Count		Sentence Count		Long Sentence Count (i)		Reading Time (mm:ss)		Dale-Chall Difficult Words (ii)		Automated Readability Index (iii)		Passive Voice Count	
Topic	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR	ChatGPT	SIR
Abdominal Aortic Aneurysm (AAA)	487	426	25	23	21	18	2:09	1:53	191	147	12.36	11.06	4	3
Cancer	873	465	39	28	32	19	3:52	2:04	339	185	13.94	11.01	12	4
Deep venous disease	979	856	52	47	42	32	4:21	3:48	356	309	10.61	10.03	9	9
Glossary of IR Terms	3801	1783	189	77	179	71	16:53	7:55	1340	680	11.31	13.77	97	28
Men's Health	988	738	46	38	39	26	4:23	3:16	421	308	13.93	12.88	15	5
Peripheral artery disease (PAD)	1275	871	68	43	52	29	5:40	3:52	438	288	10.65	12.35	17	11
Palliative care	1199	400	59	24	46	14	5:19	1:46	471	149	12.74	10.92	15	4
Stroke	1026	421	49	23	41	14	4:33	1:52	338	158	11.45	10.89	9	7
Varicose veins	845	469	53	31	42	16	3:45	2:05	356	193	10.61	10.04	21	5
Venous Disease	261	234	17	14	14	7	1:09	1:02	93	87	8.77	9.31	3	0
Women's Health	1603	1254	76	70	59	49	7:07	5:34	676	508	13.37	11.77	20	16
Grand Totals	13337	7917	673	418	567	295	59:11	35:07	5019	3012	129.74	124.03	222	92
Mean	1212.5	719.7	61.2	38	51.5	26.8	5:22	3:11	456.3	273.8	11.8	11.3	20.2	8.4

i) Long sentence count defined as sentences containing > 20 syllables. ii) Dale-Chall Difficult Words are those not found on a list of about 3,000 common words, which 80% or more of US 4th grade students are familiar with. iii) The calculation for the Automated Readability Index (ARI) is outlined in Table XXX, and assesses the approximate U.S. grade level required to read a piece of text.

Keywords

ChatGPT; Patient information; Chatbot; AI; Interventional Radiology

Scientific Abstract Presentations
Intelligent Imaging

Date:
SUN, OCT 1

Time:
10:45 AM – 12:15 PM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

Automated Detection of Pericoronary Adipose Tissue Attenuation to Detect Inflammation on Coronary Computed Tomography Angiography
+ Devina Chatterjee, Medical Student, University of Maryland School of Medicine
+ Adway Kanhere, MS; Benjamin L. Shou; Sangmita Singh; Vishwa Parekh, PhD; Paul l. Yi, MD, MS; Armin Zadeh, MD, PhD

Computer Vision-Derived Bone Mineral Density Measures of Thoracic Vertebra using Multiplanar Segmentation of Conventional Chest CT
+ Quincy A. Hathaway, MD, PhD, Resident Physician, West Virginia University School of Medicine
+ Arta Kasaeian; Elena Ghotbi, MD; Hamza Ibad, MD; João A. Lima, MD, MBA; Shadpour Demehri, MD

Deformable Multi-modal Image Registration via Neural Optimal Transport: An Application to Multi-parametric MRI Registration
+ Boah Kim, PhD, Postdoctoral Fellow, National Institutes of Health Clinical Center;
+ Tejas Mathai, PhD; Ronald M. Summers, MD, PhD

Evaluating Non-Anatomic Sequences to Predict the MGMT Status of Glioblastomas using Deep Learning
+ Emma Barry, Medical Student, University of Maryland School of Medicine
+ Vivian Zhang; Crystal Li; Peter Kamel, MD; Paul Yi, MD, MS

Learning to Generalize towards Unseen Domains via a Content-Aware Style Invariant Framework for Disease Detection from Chest X-rays
+ Taufiq Hasan, PhD, Associate Professor, Bangladesh University of Engineering and Technology
+ Mohammad Zunaed

Using AI to create Deep Non-Contrast (DNC) Images with Photon-Counting CT: A Phantom Study
+ Todd Soesbe, PhD, Assistant Professor, UT Southwestern Medical Center
+ Yee Ng, MD; Jesse Rayan, MD; Yin Xi, PhD; Dan Nguyen, PhD

Automated Detection of Pericoronary Adipose Tissue Attenuation to Detect Inflammation on Coronary Computed Tomography Angiography

Devina Chatterjee, Medical Student, University of Maryland School of Medicine; Adway Kanhere, MS; Benjamin L. Shou; Sangmita Singh; Vishwa Parekh, PhD; Paul I. Yi, MD, MS; Armin Zadeh, MD, PhD

Introduction

Pericoronary adipose tissue attenuation (PCAT), obtained from cardiac CT angiography (CCTA), has been associated with coronary inflammation and manifestations of atherosclerotic disease. However, assessment of PCAT is time-consuming and not readily available. Integration of artificial intelligence (AI) into the analysis workflow of PCAT, e.g., segmentation of arterial segments and quantification of data, may enhance the accuracy, efficiency, and reliability of PCAT assessment.

Hypothesis

Automated segmentation of perivascular fat and extraction of PCAT data on the proximal right coronary artery can be achieved using AI, which will allow differentiation between healthy patients and those with coronary artery disease.

Methods

To train the perivascular fat segmentation model on the coronary arteries, we used 70 3D CTA images from the ImageCAS dataset, with 70 scans for internal validation. Using the nnUnet deep learning model, we segmented the proximal (10-50 mm away from the opening) right coronary artery (RCA) and extracted the perivascular fat region (-190 to -30 HU) around the coronary arteries up to 5 cm away from the vessel wall. The external validation dataset, ASOCA, included CTAs and segmentations of the RCA for 20 healthy patients and 20 patients with confirmed coronary artery disease. Manual extraction of PCAT was performed on the external validation dataset was performed using the software 3D-slicer.

Results

Automated segmentation of the proximal RCA yielded a DICE score of 0.896 and 0.822 for the internal and external validation cohorts, respectively. The mean attenuation values for PCAT at 5 mm in the healthy and diseased patients is -76.3 ± 5.46 HU and -72.1 ± 4.82 HU respectively. Compared to the manual measurements, for the healthy and diseased patients, the intraclass correlation coefficient is 0.916.

Conclusion

In this study, we demonstrated the feasibility of automated segmentation of perivascular fat and extraction of PCAT using AI. Our results showed high accuracy and reliability of the AI model in segmenting the proximal RCA and quantifying PCAT. The findings support the hypothesis that AI-based analysis can differentiate between healthy individuals and those with coronary artery disease based on PCAT characteristics.

Statement of Impact

The localization and severity of inflammation in proximity to atherosclerotic plaques may provide valuable information regarding the disease activity and associated risk of acute events. This research highlights the potential of AI in improving PCAT analysis and its significance in cardiovascular disease diagnosis and treatment. Future work should focus on validating these findings on larger datasets and utilizing AI to extract more 3-D radiomic features from inflammation.

Figure 1: PCAT was automatically detected with the deep learning model in axial CCTA scans in both the control group (A-D) and in patients with confirmed CAD (E-G). In (B) and (F), the automated detection of the coronary artery and the perivascular region 10 mm around the vessel wall is highlighted in red. In (C) and (G), the perivascular fat region defined by HU values from -190 to -30 is highlighted in red up to 10 mm away from the vessel wall. In (D) and (H) the zoom in highlights the inflammation in the perivascular fat. A heat map is overlayed the region where yellow indicates lower levels of inflammation (closer to -190 HU) and red indicates higher levels of inflammation (closer to -30 HU)

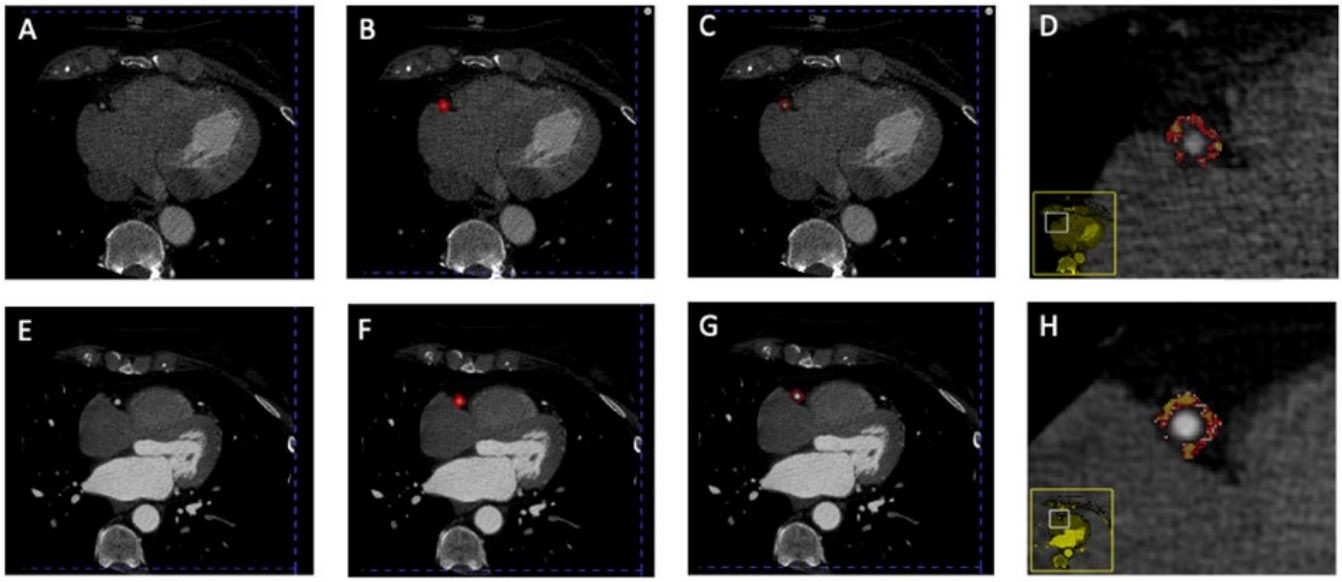


Figure 2: (A) shows the median attenuation of the perivascular fat around the coronary arteries. All shells located radially outward had $p < 0.05$ except for shells 1 and 2 mm out from the coronary artery. (B) represents a ROC curve predicting normal vs disease in the external validation cohort with the median attenuation of perivascular fat. (C) highlights the significant difference of the median attention on the perivascular fat located 5 mm out from the vessel wall.

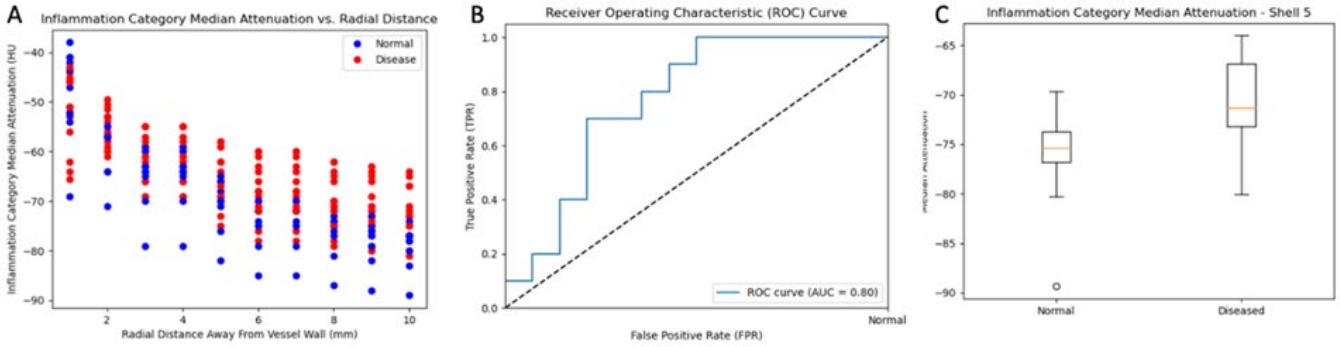
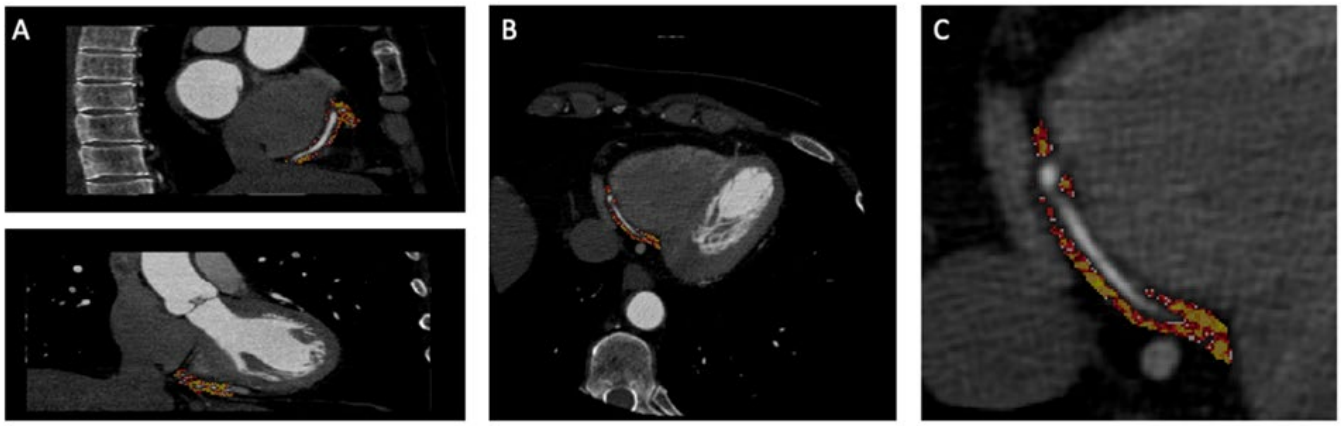


Figure 3: Extension of the PCAT segmentation can be applied beyond the proximal RCA. Located ~60 mm away from the opening of the RCA, the inflammatory status of a potential legion is analyzed. (A) shows the coronal and sagittal views of the lesion and (B) shows the same lesion in the axial view. (C) shows the heat map is overlayed the region where yellow indicates lower levels of inflammation (closer to -190 HU) and red indicates higher levels of inflammation (closer to -30 HU)



Keywords
Pericoronary adipose tissue; Coronary artery disease; Atherosclerosis; Inflammation; Artificial intelligence

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Computer Vision-Derived Bone Mineral Density Measures of Thoracic Vertebra using Multiplanar Segmentation of Conventional Chest CT

Quincy A. Hathaway, MD, PhD, Resident Physician, West Virginia University School of Medicine; Arta Kasaeian; Elena Ghotbi, MD; Hamza Ibad, MD; João A. Lima, MD, MBA; Shadpour Demehri, MD

Introduction
While dual x-ray absorptiometry (DEXA) is considered the gold standard for measuring bone mineral density (BMD), quantitative computed tomography (QCT) has been rapidly emerging for volumetric BMD (vBMD) measurements even in the absence of calibrating phantoms (i.e., phantom-less QCT). QCT offers the ability to measure various orientations, perform vBMD determination at multiple vertebral levels, and potentially improve fracture risk prediction. No current QCT-based automated segmentation approaches capture information across multiple vertebral segments and orientations.

Hypothesis
We hypothesized that consecutive thoracic vertebrae in both axial and sagittal frames can be accurately automated in QCT.

Methods
3,083 non-contrast, chest CTs were obtained. An automated platform was developed to obtain the primary axial sequences and convert DICOM images to PNGs with a W:1800 and L:400. Using pixel-based intensities of the vertebral bone, consecutive axial images from T1-T10 were obtained (Figure 1). Likewise, pixel-based intensities were used to automate selection of sagittal frames for the upper and lower spine. Manual segmentation of 600 patients (n=360 training, n=240 testing) for the axial (n=6,000 images) and sagittal (n=1,200 images) frames using LabelMe (v5.0.3) was performed. PixelLib (v0.7.1) and Mask R-CNN (v2.1) were employed to create a custom instance segmentation platform, with transfer learning through ResNet50. Dice coefficient was calculated.

Results
The mean Dice coefficient for the axial images was 0.94 (95%CI: 0.95–0.94) training and 0.93 (95%CI: 0.94–0.92) testing. The mean Dice coefficient for the sagittal images was 0.89 (95%CI: 0.91–0.87) training and 0.87 (95%CI: 0.88–0.85) testing. A total of 30,083 axial and 6,166 sagittal images were segmented (Figure 2). Mean Hounsfield units and radiomics were collected from the regions of interest.

Conclusion
Our automated approach provides a feasible option for multiplanar, contiguous thoracic spine assessment. Through this robust automated methodology, variability in spine bone mineral density, radiomics, and topology data can be easily acquired and help augment prediction of fracture risk.

Statement of Impact
Automation of sagittal and axial spine segmentation is the first step in personalizing fracture risk profiles for patients by supplementing traditional bone mineral density measurements with detailed information of the vertebral body architecture.

Figure 1: Pixel intensity-based frame selection. Axial images were filtered based on pixels with a 0-255 greyscale value greater than or equal to 90. The histogram represents the total pixels in the vertebrae, ranging from T1-T10. In-between the peaks and the troughs represent an approximate depth of 5 mm within the vertebrae.

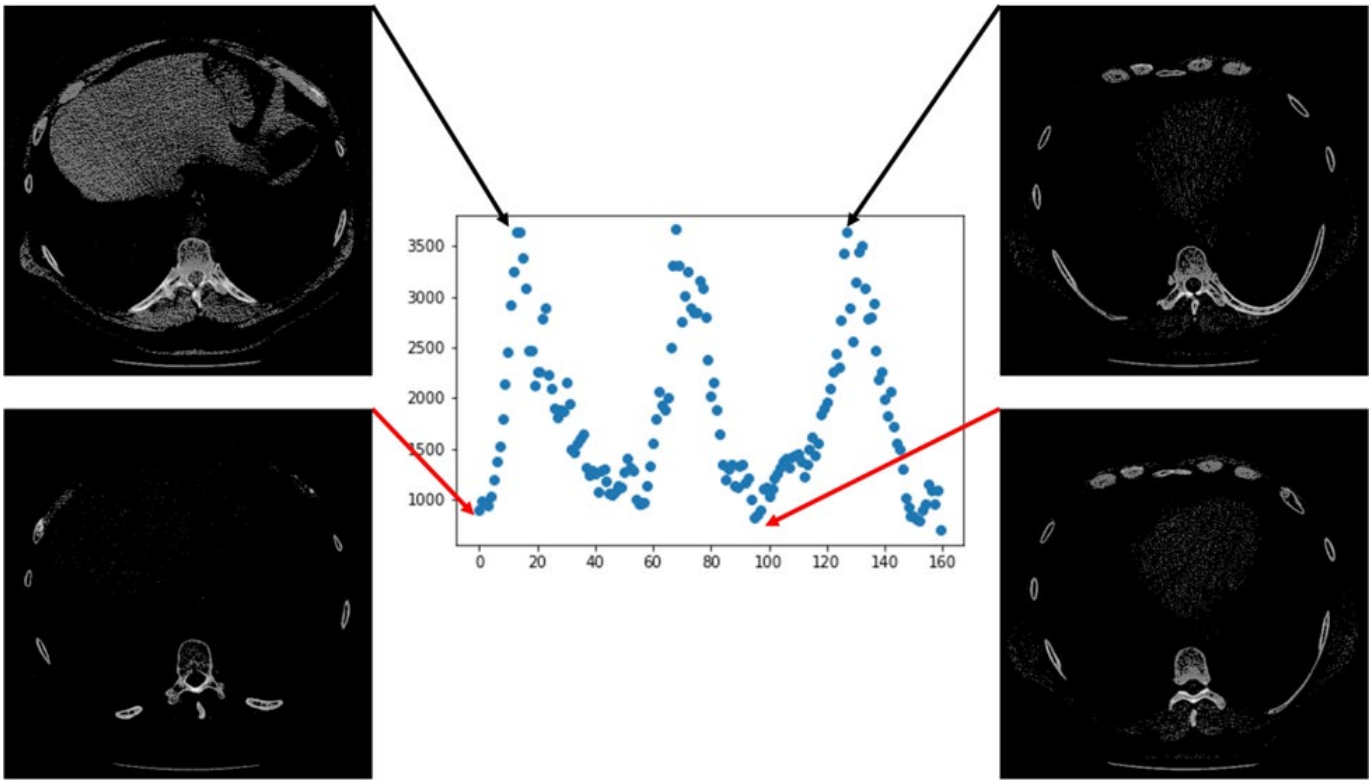
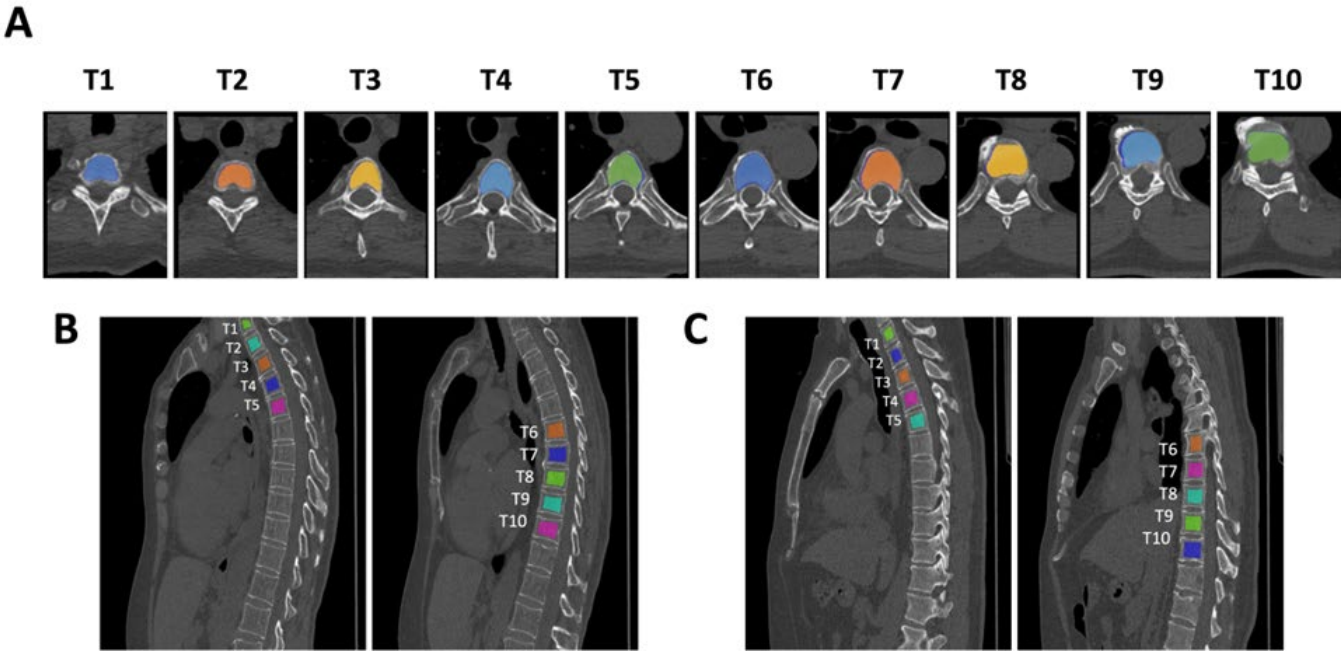


Figure 2: Instance segmentation in axial and sagittal images. (A) Example of axial segmentation from T1 (left) to T10 (right). ROI was designed to not include the cortical bone. (B) Example of sagittal segmentation of T1-T5 (left) and T6-T10 (right). (C) Example of how using two separate groups of ROIs helps to better capture patients with scoliosis.



Keywords
Instance Segmentation; Bone Mineral Density; Transfer Learning; Spine; Vertebral Body

Deformable Multi-modal Image Registration via Neural Optimal Transport: An Application to Multi-parametric MRI Registration

Boah Kim, PhD, Postdoctoral Fellow, National Institutes of Health Clinical Center; Tejas Mathai, PhD; Ronald M. Summers, MD, PhD

Background/Problem to be solved

Multi-parametric MRI sequences are often used for diagnosing diseases, such as lymphoma. However, since there are volumetric misalignments due to scan time, patient breathing, etc., deformable image registration is essential to analyze multi-modal data at the same coordinates. Although recent unsupervised learning-based methods are widely applied to various registration tasks, multi-modal image registration is a still challenging problem in that different modality images have different data distributions. To address this, we propose a deep learning framework for multi-modal registration.

Intervention(s)

We design an unsupervised domain-transported image registration model, named OTMorph, by employing a recent neural optimal transport approach. Our model is composed of two modules: a transport module translates the data distribution from the moving source domain to the fixed target domain, and a registration module provides deformation for the moving image to be aligned into the fixed image. Accordingly, when a pair of moving and fixed images is given, the transport module learns deterministic optimal transport maps between the inputs. Then, the registration module takes the translated map of the moving image and outputs the deformation fields to warp the moving image. Through end-to-end learning, the proposed model can effectively learn deformable registration for the images in different distributions.

Outcome

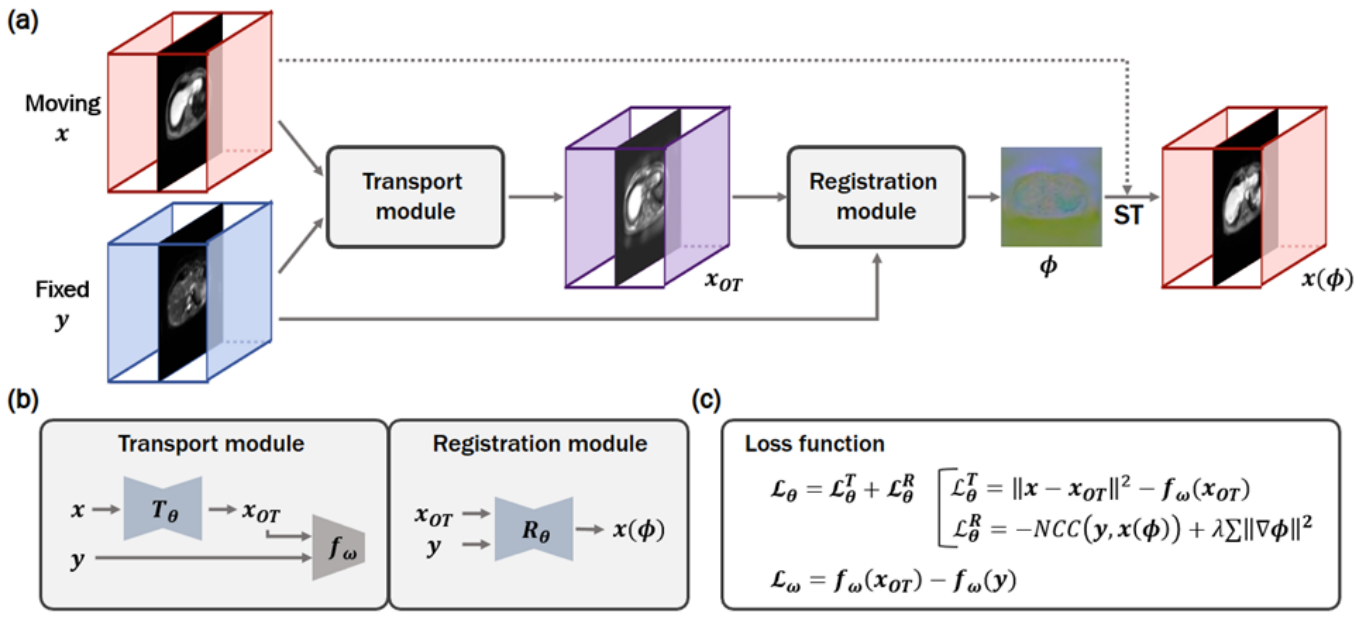
The proposed method is demonstrated on abdominal MRI registration from T1-weighted (T1w), T2, and fat-suppressed T2 (T2FS) imaging to diffusion-weighted imaging (DWI). When training our model with a diffeomorphic layer, ours improves the accuracy of T1w-to-DWI registration with a 17% gain in Dice score on the segmentation maps of several organs (0.70 vs. 0.53) over the initial resampled data. Compared to the baseline methods, ours shows a 4% improvement (0.70 vs. 0.66). The T2/T2FS-to-DWI registration performance of ours also shows comparable results with the others.

Conclusion

We present a learning-based multi-modal image registration model using the neural optimal transport algorithm. Our model computes an optimal transport plan to translate the moving image to the fixed image data distribution, and then instead of the original moving image, the registration module uses the transported image, enabling the network to estimate deformation fields more effectively. Experimental results on multi-modal MRI registration show that the proposed method is superior for deformable registration compared to the existing learning-based methods.

Statement of Impact

Our proposed framework is generic and can be used for intra-/inter-modality registration tasks, especially when the underlying intensity distributions are vastly different. We believe that this facilitates many downstream clinical goals, such as follow-up tracking of tumors.



(a) The overall framework of the proposed method. For a given input pair of moving image x and fixed image y, the transport module estimates x_OT that is transported from the moving to fixed data distribution, then the registration module yields the deformed image x(ϕ) that is warped using the deformation field ϕ through a spatial transformation layer (ST). (b) The detailed training flow of the proposed method, where T and f are the optimal transport mapping network and potential network, respectively, and R is the registration network. θ and ω are network parameters. (c) The loss function for training the model, where NCC is a normalized cross-correlation function.

(a)

	Train	Validation	Test
The number of patients	1363	170	170
Manufacturer	Siemens		
Data processing	1. Resample T1w/T2/T2FS to have the same voxel size with DWI 2. Resize the data into 256×256×48		

(b)

Method	T1w → DWI	T2 → DWI	T2FS → DWI
Initial (Resampled)	0.532 (0.251)	0.655 (0.275)	0.817 (0.221)
Affine	0.647 (0.250)	0.636 (0.279)	0.828 (0.217)
VoxelMorph	0.660 (0.248)	0.657 (0.271)	0.844 (0.221)
VoxelMorph-diff	0.664 (0.247)	**0.670 (0.277)**	0.843 (0.221)
CycleMorph	0.669 (0.248)	0.664 (0.274)	0.846 (0.222)
OTMorph	0.684 (0.244)	0.665 (0.274)	**0.849 (0.222)**
OTMorph-diff	**0.701 (0.250)**	0.663 (0.274)	0.846 (0.222)

*Average Dice scores for liver, spleen, left and right kidneys

(a) Information of the dataset we used. (b) Quantitative evaluation results of multi-parametric MR image registration from Affine by ANTsPy library of Python (Avants, Brian B., et al., Insight Journal, 2009), VoxelMorph (Balakrishnan, Guha, et al., CVPR 2018), VoxelMorph-diff (Dalca, Adrian V., et al., MICCAI 2018), CycleMorph (Kim, Boah, et al., Medical Image Analysis 2021). OTMorph-diff is the proposed method with the diffeomorphic layer. We warp the

segmentation maps of the moving image using the deformation fields estimated from each method and compute the average Dice coefficient between the deformed segmentation masks and the fixed image segmentation masks. Standard deviations are in parenthesis. The best scores are bolded, and the second-best scores are underlined.

Qualitative image registration results from the proposed method and several existing learning-based methods. We deform each T1w, T2, and T2FS moving image into the fixed DWI image. For each registration, the top row shows the visual results, and the bottom row shows the boundaries of deformed segmentation maps (dark/light red: liver, and dark/light green: spleen). The orange arrows indicate remarkable parts.

Keywords

Deformable image registration; multi-modal medical imaging; optimal transport; unsupervised learning

Evaluating Non-Anatomic Sequences to Predict the MGMT Status of Glioblastomas using Deep Learning

Emma Barry, Medical Student, University of Maryland School of Medicine; Vivian Zhang; Crystal Li; Peter Kamel, MD; Paul Yi, MD, MS

Introduction

MGMT promoter methylation status is used as a prognostic factor and target for chemotherapy in patients diagnosed with Glioblastoma Multiforme (GBM). Brain biopsy remains first-line in determining MGMT promoter methylation status. Deep learning has led to many attempts in using anatomic sequences such as T1 post-contrast sequences to predict the methylation status of tumors, though has been met with limited success. Functional imaging parameters which assess the diffusivity of water molecules in tissues have been associated with MGMT prediction, but a comparison of its predictivity against other MRI sequences and utility in deep learning algorithms is still unknown. We compared the performance of deep learning models trained on anatomic MRI sequences consisting of T1 pre and post-contrast, T2, T2-FLAIR to diffusion sequences of Diffusion Weighted Imaging (DWI), Apparent Diffusion Coefficient (ADC), and Fractional Anisotropy (FA) to predict MGMT methylation status.

Hypothesis

DWI-ADC sequences will outperform all other MRI sequences in MGMT prediction because of the inclusion of functional features that have been pathologically known to distinguish MGMT status of GBMs.

Methods

MRI images of pre-operative glioblastomas from the University of California San Francisco (UCSF-PDGM) were split into train and validation sets (80:20). We trained a DenseNet model architecture using the open source framework MONAI. The model was trained individually on all anatomic and diffusion-based MRI sequences to classify the MGMT promoter methylation status of the tumor. Performance was assessed using area under the curve (AUC) and compared for each sequence using the DeLong method.

Results

For anatomic sequences, models trained on the T1 post-contrast images demonstrated the highest performance (AUC 0.613), followed by T1 pre-contrast images (AUC 0.590), T2 (AUC 0.562) and T2-FLAIR images (AUC 0.540). Models trained on functional diffusion sequences significantly outperformed those trained on anatomic sequences with ADC achieving an AUC of 0.905 and DTI-FA images an AUC of 0.81 ($p < 0.001$ compared to T1 post contrast images). There was no statistically significant difference between DWI (AUC 0.643) and T1 post contrast images ($p = 0.231$).

Conclusion

The highest performance of predicting MGMT promoter methylation status was achieved by the ADC sequence,

which significantly outperformed anatomic sequences. This highlights the importance and significance of quantitative functional parameters in predicting MGMT status using deep learning.

Statement of Impact

Functional diffusion-based MRI sequences provide significant information to deep learning algorithms to predict MGMT promoter methylation status, outperforming traditional anatomic sequences. This can lead to more accurate noninvasive methods of predicting tumor biomarkers without brain biopsy.

Figure 1. Flowchart demonstrates the experiment run on the 2D DenseNet Model for training and testing. The model was trained individually on T1 pre and post-contrast, T2, T2-FLAIR, DWI, DTI-FA, and ADC sequences with a bounding box restricted to the necrotic core of the tumor. Data augmentation and cross-fold validation was performed for each of the models.

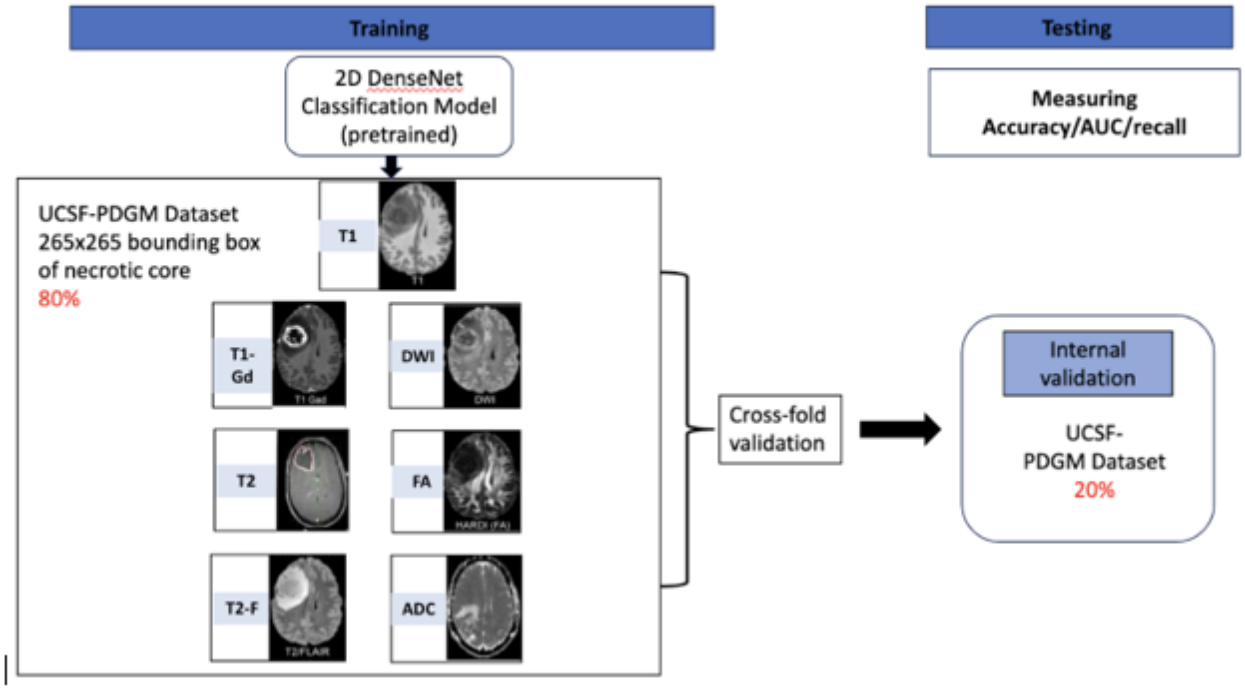
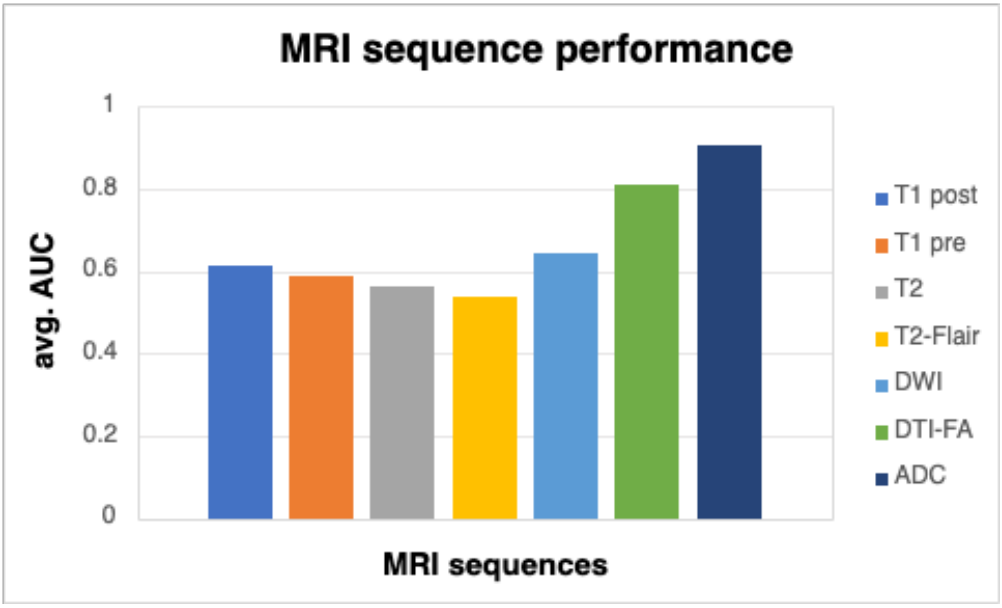


Figure 2. Average AUC obtained by each deep learning model trained on an individual MRI sequence. Models trained on the functional diffusion sequences ADC and DTI-FA significantly outperformed those trained on anatomic sequences.



Keywords
GBM; MRI; MGMT; DWI; ADC; Deep learning

Learning to Generalize towards Unseen Domains via a Content-Aware Style Invariant Framework for Disease Detection from Chest X-rays

Mohammad Zunaed and Taufiq Hasan (PhD), Department of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh.

Introduction

Machine learning techniques for medical image analysis (e.g., chest X-ray (CXR) based diagnosis) usually suffer from data distribution discrepancies, known as domain shift (DS). Domain adaptation and generalization methods, such as adversarial learning or multi-domain mixups, have been proposed to address DS on CXR datasets. However, they do not explicitly consider the generated domain-invariant features' content and style attributes. Recent literature studies have showcased that ImageNet-trained CNNs are biased toward styles (i.e., uninformative textures) rather than content.

Hypothesis

Unseen domains do not affect radiologists as they diagnose based on visual cues rather than uninformative textures. Following this, deep learning models should derive high-level features for pathology classification that are not only domain-agnostic but also content-focused and style-agnostic.

Methods

Existing domain generalization methods for natural images based on neural style transfer have some shortcomings. During training, they randomize styles to simulate unseen domains based on available source domains. However, we argue that these synthesized styles are limited to available source domains. Moreover, they utilize channel-wise mean and standard deviations as style parameters in pre-defined manners. Learnable pixel-wise style embeddings may result in more diverse stylizations. To address these shortcomings, we employ an image-level style randomization module (SRM) that randomly samples style parameters from a set constructed by the prior possible domain value range of a CXR image compared to utilizing the existing source domains. Furthermore, we use a learnable feature-level SRM that randomizes the style attributes on the pixel level. In addition, we leverage two regularization losses, named content consistency regularization and probability distribution regularization, between a CXR image and a style-perturbed version of that CXR image to tweak the framework's sensitivity toward content markers.

Results

We conduct experiments in the cross-domain setting by training on the CheXpert and MIMIC-CXR datasets and evaluating on the BRAX dataset for the thoracic disease classification task. Our proposed method achieves 77.36±0.09 mean AUC compared to 75.90±0.36 from the former best model (i.e., 1.92% relative improvement) on a cross-validation setup with statistically significant results.

Conclusion

Experiments under the cross-domain setting demonstrate that a model that is style-agnostic but biased toward content-specific features, i.e., pathology-specific cues, is more robust in the presence of domain shifts compared to the state-of-the-art approaches.

Statement of Impact

This work establishes a custom-designed model framework and training approach to manipulate styles and exploit content consistency to yield a domain-agnostic, style-invariant, and content-biased model that performs well on unseen domains and is more akin to real-life radiologists.

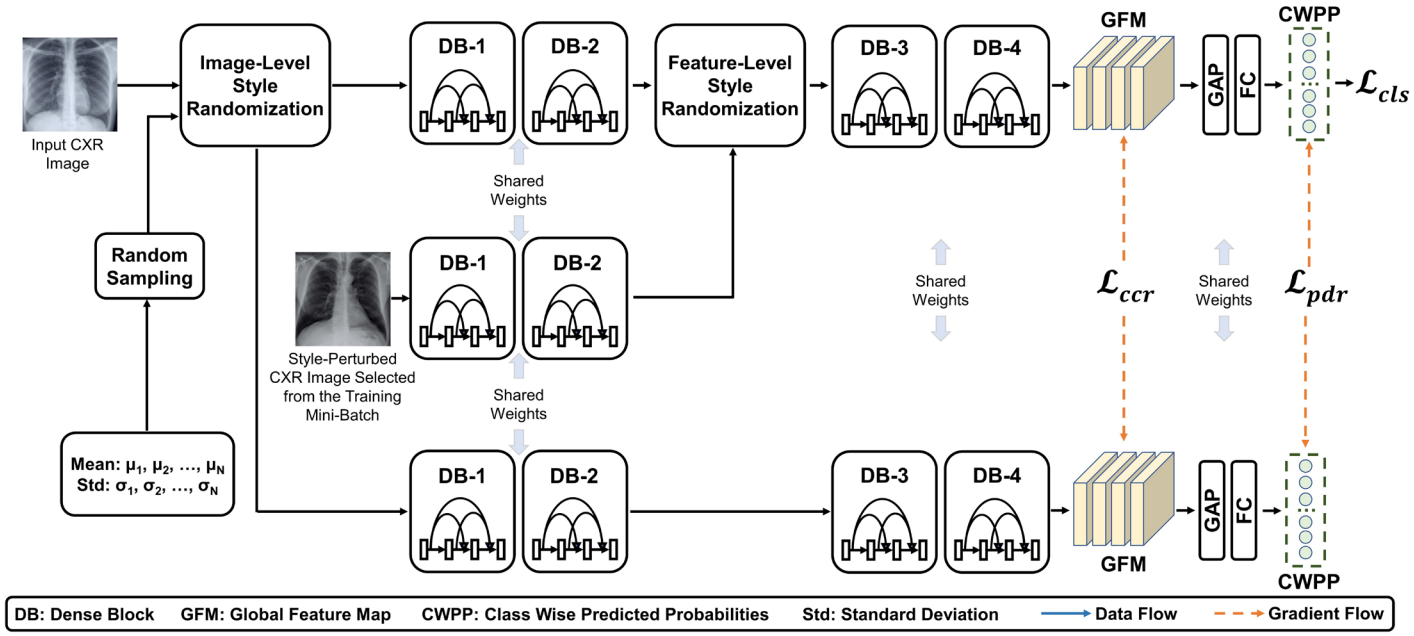


Figure 1: Illustration of the proposed framework. The input CXR image is perturbed with randomly sampled style statistics (i.e., mean and standard deviation) from a set of values defined by prior domain knowledge. The style statistics of the image-level style-perturbed CXR image are further augmented at the feature level with another stylized CXR image's feature-level style statistics selected from the training mini-batch. Two consistency regularization losses, i.e., content consistency loss ($\mathcal{L}_{ccr}$) and predicted probability distribution loss ($\mathcal{L}_{pdr}$), are applied between the global feature spaces and predicted probability distributions, respectively, for the same CXR image with and without style statistics perturbed. The image-level and feature-level style-perturbed CXR image's global feature map is pooled and utilized for pathology classification ($\mathcal{L}_{cls}$).

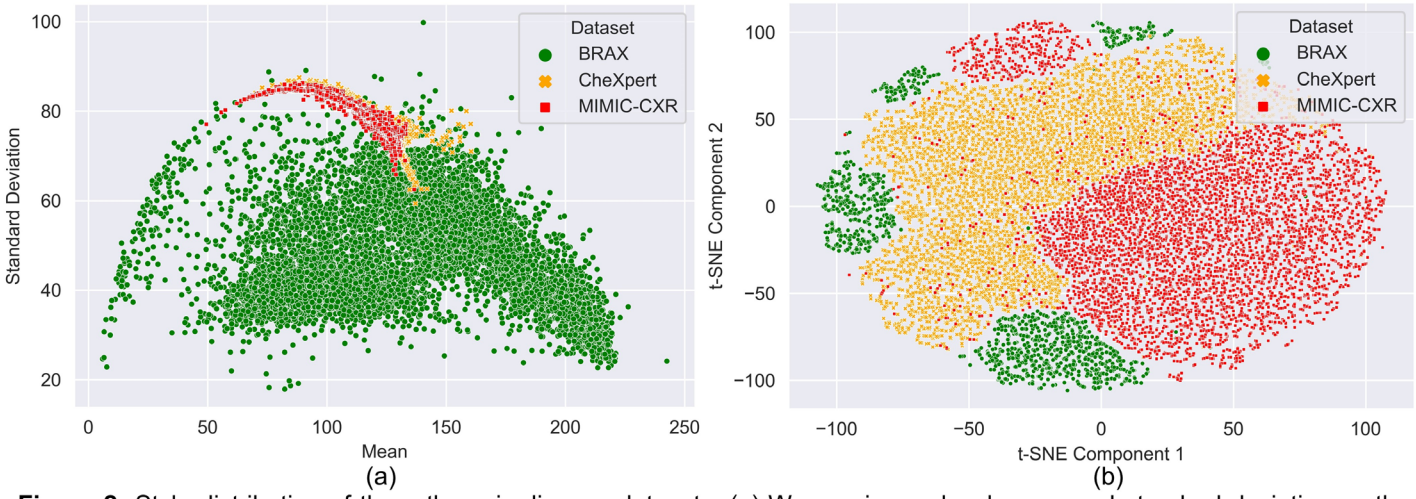


Figure 2: Style distribution of three thoracic disease datasets. (a) We use image-level mean and standard deviation as the style feature and observe a style distribution gap between different datasets. (b) Feature-level style statistics visualization (2D t-SNE components of the concatenation of means and standard deviations, computed from the convolutional feature maps from the DenseNet-121 CNN architecture's first dense block, trained with the three thoracic datasets, i.e., CheXpert, MIMIC-CXR, and BRAX. The feature statistics clearly capture dataset-specific uninformative information, i.e., styles observed by the separable groups.

Table 1
Comparison with state-of-the-art approaches on the unseen domain test dataset (BRAX dataset[†]).

Method	Atel	Card	E.C.	Cons	Edem	Pne1	Pne2	P.E.	P.O.	L.L.	L.O.	Frac	S.D.	N.F.	Average	p-value[‡]
Luo *et al.* (Luo et al., 2020)	74.92	87.14	66.25	64.03	75.17	87.30	79.09	85.61	83.52	57.55	71.47	61.43	76.96	70.24	74.33±0.80	0.0094*
CrossNorm & SelfNorm (Tang et al., 2021)	74.76	86.82	66.10	67.17	75.15	86.14	78.61	84.82	84.17	58.20	70.57	61.12	77.43	70.01	74.36±0.43	0.0017*
FDA (Yang and Soatto, 2020)	76.60	84.93	60.98	70.27	75.31	87.45	78.77	86.47	87.05	55.81	71.84	61.83	77.52	70.45	74.66±0.27	0.0010*
pAdaIN (Nuriel et al., 2021)	75.30	87.03	65.82	67.72	74.84	87.34	77.06	85.91	87.34	58.95	71.55	62.17	76.99	70.66	74.91±0.29	0.0011*
SagNet (Nam et al., 2021)	75.48	87.73	66.77	67.49	74.26	87.00	78.49	86.01	88.93	58.98	71.36	61.86	77.06	70.76	75.15±0.37	0.0031*
AdvStyle (Zhong et al., 2022)	75.31	87.03	67.01	68.17	75.53	87.85	77.57	86.08	88.45	58.16	71.52	62.40	76.81	70.53	75.17±0.23	0.0010*
STRAP (Yamashita et al., 2021)	74.54	86.40	65.11	70.52	76.80	87.21	78.45	86.74	91.47	55.85	70.64	61.83	76.91	71.06	75.25±0.32	0.0009*
Wang *et al.* (Wang and Xia, 2023)	75.23	86.11	66.16	69.41	75.37	87.75	79.19	86.22	89.67	56.88	71.35	63.47	76.89	70.40	75.29±0.32	0.0020*
MixStyle (Zhou et al., 2021a)	75.62	88.05	66.53	67.75	75.33	87.27	78.57	86.48	89.78	59.03	71.96	62.45	77.88	71.21	75.57±0.16	0.0008*
FSR (Wang et al., 2022)	75.92	87.47	69.12	69.97	75.87	87.91	77.17	86.05	88.59	59.23	72.97	63.46	77.37	71.51	75.90±0.36	0.0060*
Ours	77.82	89.05	71.01	67.92	75.07	87.95	79.40	89.35	93.13	61.37	73.21	68.20	77.19	72.36	77.36±0.09	REF

[†] The 14 findings are Atelectasis (Atel), Cardiomegaly (Card), Enlarged Cardiomediastinum (E.C.), Consolidation (Cons), Edema (Edem), Pneumonia (Pne1), Pneumothorax (Pne2), Pleural Effusion (P. E.), Pleural Other (P.O.), Lung Lesion (L.L.), Lung Opacity (L.O.), Fracture (Frac), Support Devices (S.D.), No Finding (N.F.).
[‡] Statistical significance compared to the proposed method (REF) is measured utilizing the paired t-test on cross-validated mean AUC scores.
[*] Indicates a significant difference.

Keywords
Chest X-ray, Domain shift, Feature statistics, Style randomization, Thoracic disease classification

Using AI to create Deep Non-Contrast (DNC) Images with Photon-Counting CT: A Phantom Study

Todd Soesbe, PhD, Assistant Professor, UT Southwestern Medical Center; Yee Ng, MD; Jesse Rayan, MD; Yin Xi, PhD; Dan Nguyen, PhD

Introduction

Both photon-counting CT and dual-energy CT scanners offer virtual non-contrast (VNC) images where the X-ray attenuation from intravascular iodine is virtually removed. VNC images help differentiate iodine enhancing tissues from other hyperattenuating materials like calcium, which is beneficial for post-contrast imaging of coronary calcium and kidney stones. However, current VNC methods affect all pixels within the CT image, and not just those that contain iodine. These unwanted VNC artifacts can significantly alter the HU values for calcium and fat, leading to inaccuracies in coronary calcium scores and kidney stone characterization.

Hypothesis

Deep learning can use the multi-energy information from spectral CT scanners to create deep non-contrast (DNC) images where only the iodine attenuation is removed, leaving all other HU values unchanged.

Methods

A multi-energy CT phantom (Sun-Nuclear, 300x400-mm) was configured with nine tissue-equivalent rods (iodine at 2, 5, 10, 15, 20-mg/ml, calcium at 100, 200, 300-mg/ml, and fat) and one space for air. The phantom was scanned with photon-counting CT (Naeotom Alpha, Siemens) using an Abdominal/Pelvis protocol (144x0.40-mm, 140-kV, 14-mGy). Monoenergetic (40, 70, 190-keV) and VNC images were reconstructed using Siemens software (256x256-mm FOV, twenty 5-mm thick slices, Br44, QIR-3). The five iodine rods were then replaced with solid-water rods and the phantom re-scanned using the same protocol. This two-scan pair was repeated for 25 randomized rod configurations, giving 500 examples. The 40, 70, and 190-keV iodine rod images were inputs into a U-Net (5-levels, ReLU, 38.4M parameters, Adam optimizer, MSE loss) with the 70-keV water rod image as the true non-contrast (TNC) ground truth. The model was trained on for 500 epochs using a 360/40/100 split for training/validation/testing (i.e., 18/2/5 unique scans). HU biases for VNC and DNC images were measured using circular ROI (3.0-cm^2) placed over each rod with TNC as the reference. Overall accuracy between the TNC, VNC, and DNC images was calculated using root mean squared error (RMSE).

Results

Biases from DNC were smaller (-2.9, -16.2, -0.6 HU) than VNC (-12.5, -381.1, +11.9 HU) for all iodine-to-water, calcium, and fat ROIs, respectively. Overall RMSE was also smaller for DNC (0.007816) than VNC (0.035548) when using TNC as the reference.

Conclusion

Deep learning can use the muti-energy information from photon-counting CT scanners to remove iodine attenuation while leaving all other tissue HU values unchanged. This method can apply to all spectral CT scanners.

Statement of Impact

DNC images could increase spectral CT diagnostic accuracy and eliminate the need for pre-contrast scans in multi-phase studies.

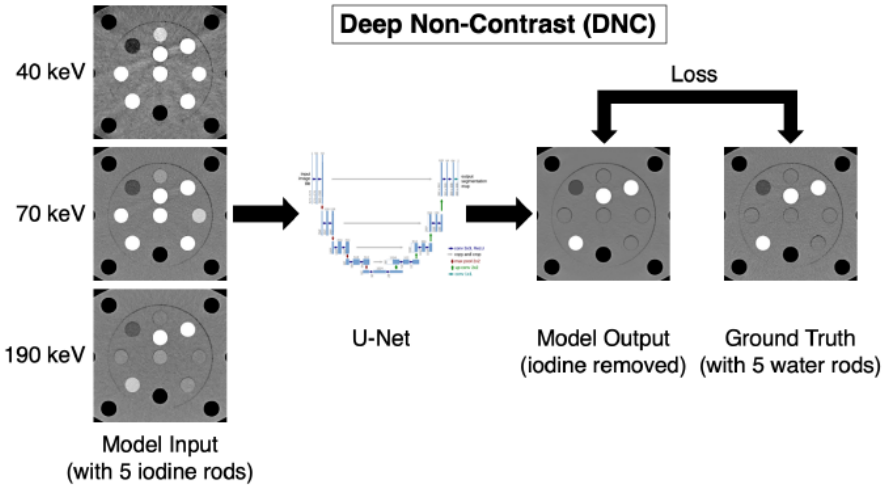


Figure 1: Schematic of the Deep Non-Contrast (DNC) model training. A three-channel tensor containing the 40, 70, and 190-keV monoenergetic images (with the 5 iodine rods) was used as input into the U-Net model. The model output was then compared to the 70-keV ground truth image (where the iodine rods were replaced with water rods). Mean squared error (MSE) loss was then calculated and minimized over the 500 training epochs.

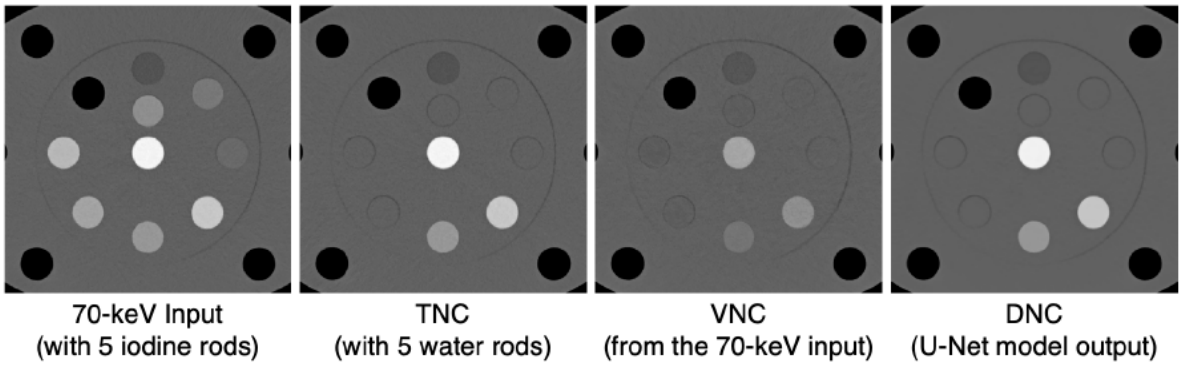


Figure 2: The 70-keV input, true non-contrast (TNC) ground truth, virtual non-contrast (VNC), and deep non-contrast (DNC) images for a single testing example. Note that the attenuation of the three calcium rods is significantly reduced in the VNC image compared to the DNC image. Window level/width = 300/1500 HU for all.

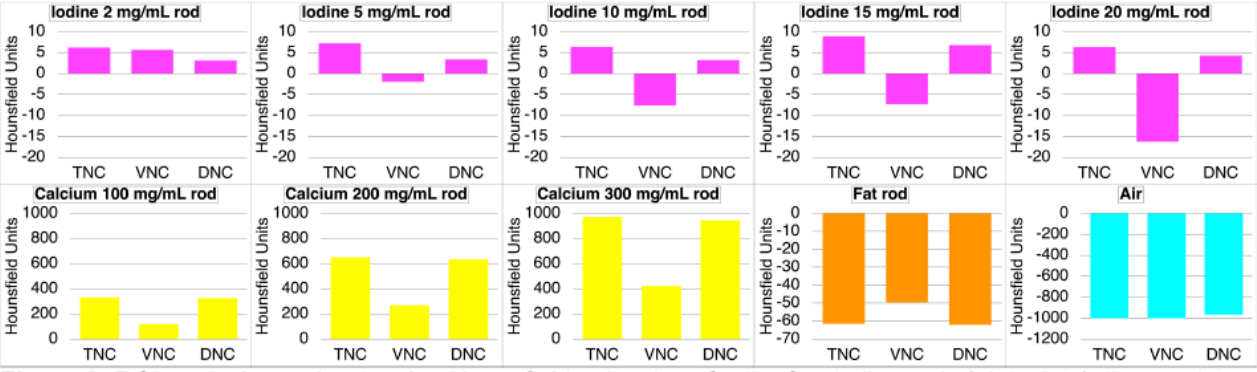


Figure 3: ROI analysis results showing Hounsfield unit values for the five iodine rods (pink plots), three calcium rods (yellow plots), fat rod (orange plot), and air (blue plot) within the true non-contrast (TNC), virtual non-contrast (VNC), and deep non-contrast (DNC) images. ROI values were averaged across each image type within the testing dataset. Averaged biases from DNC were smaller (-2.9, -16.2, -0.6 HU) than VNC (-12.5, -381.1, +11.9 HU) for all iodine-to-water, calcium, and fat ROIs, respectively.

Keywords

Photon-Counting CT; Virtual Non-Contrast; Deep Learning; U-Net; CT Phantom; Spectral CT

Scientific Abstract Presentations
Data Sets, Emerging Technologies & NLP Models

Date:
SUN, OCT 1

Time:
2:45 PM – 4:15 PM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

Attention-Based Weakly Supervised Deep Learning Model for Predicting the Progression of Barret's Esophagus to HGD/EAC using Pre-progression Whole Slide Images
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic Rochester
+ Chamil Codipilly, MD; Mana Moassefi, MD; Bardia Khosravi, MD, MPH, MHPE; Pouria Rouzrokh, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Prasad Iyer, MD

ConTEXTual Net: A Multimodal Vision-Language Model for Segmentation of Pneumothorax
+ Zachary Huemann, Graduate Student Researcher, University of Wisconsin-Madison
+ Xin Tie; Junjie Hu, PhD; Tyler J. Bradshaw, PhD

The Brain Tumor Segmentation (BraTS-METS) Challenge 2023
+ Ahmed W. Moawad, MD, Radiology Resident, Mercy Catholic Medical Center
+ Anastasia Janas, MD, PhD; Nourel Tahoon, MD; Spyridon Bakas, PhD; Mariam S. Aboian, MD, PhD

Using An Open-source Language Model to Abstract the Presence of Acute Cervical Spine Fracture from Radiologic Reports: A HIPAA Compliant Alternative to "ChatGPT"
+ Bardia Khosravi, MD, MPH, MHPE, Postdoctoral Research Fellow, Mayo Clinic AI Lab
+ Sanaz Vahdati, MD; Pouria Rouzrokh, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Ali Ganjizadeh, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Using Graph Representation Learning to Passively Learn Imaging Protocols
+ Dimitri Falco, PhD, AI Engineer, Quantivly
+ Robert MacDougall, MS; Benoit Scherrer, PhD

WESTERN-RLP: Augmenting Image-Caption Radiology Datasets using Image Embeddings Search of Large-Scale Natural Image Databases
+ Kartik Gupta, Medical Student, University of Western Ontario
+ David Li, MD; Jaron Chong, MD

Attention-Based Weakly Supervised Deep Learning Model for Predicting the Progression of Barret's Esophagus to HGD/EAC using Pre-progression Whole Slide Images

Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic Rochester; Chamil Codipilly, MD; Mana Moassefi, MD; Bardia Khosravi, MD, MPH, MHPE; Pouria Rouzrokh, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Prasad Iyer, MD

Introduction

The incidence of esophageal adenocarcinoma is on the rise, particularly in Western societies, and it represents the most prevalent subtype of esophageal cancer. The development of dysplasia in Barrett's esophagus (BE) is a key step in the progression to esophageal adenocarcinoma. However, accurately identifying the risk of progression in BE patients is a significant challenge, as the current method of using dysplasia grade for patient management has limitations. The diagnosis of dysplasia can vary widely among even expert pathologists, leading to a 20-fold difference in annual progression rates for low-grade dysplasia (LGD) cohorts (ranging from 0.5% to over 10%). This variation complicates the process of making recommendations for patient management. In this study, we aimed to develop a weakly supervised deep learning (DL) model using multiple instance learning (MIL) with an attention-based architecture to predict progressive cases based on pre-progression whole slide images (WSIs).

Hypothesis

Without the need for patch-level annotations, attention-based weakly supervised DL models have the capability to identify patches that serve as indicators for BE progression, solely relying on the label provided at the WSI level.

Methods

To conduct our study, we gathered baseline pre-progression WSIs of LGD cases with progression to high-grade dysplasia (HGD) or esophageal adenocarcinoma (EAC), as well as LGD cases without progression over a period of five years, from our institution. Previously, we trained a DL model to predict different levels of dysplasia in WSIs (Faghani et al., 2022). In this study, we applied our model to WSIs to identify regions of interest corresponding to non-dysplastic Barrett's esophagus (NDBE) and LGD. Each WSI was treated as a collection of instances, comprising either NDBE or LGD patches, in our MIL pipeline. Using a 10-fold cross-validation approach, we trained an attention-based MIL model. We also generated an attention map for visualizing the model's focus.

Results

Our study included 27 LGD cases with progression and 44 LGD cases without progression. The developed DL prognostic model achieved a mean area under the receiver operating characteristic curve (AUROC) of 0.85, with a standard deviation of 0.04.

Conclusion

In conclusion, we successfully developed and internally validated a DL prognostic model capable of predicting the progression of BE to HGD or EAC after five years from the initial diagnosis.

Statement of Impact

The development of a DL model for predicting the progression of Barrett's esophagus has the potential to revolutionize patient management by providing accurate prognostic information, thus improving outcomes and enabling timely interventions.

Keywords

Deep learning; Multiple instance learning; Weakly supervised learning; WSI; Barrett’s esophagus

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

ConTEXTual Net: A Multimodal Vision-Language Model for Segmentation of Pneumothorax

Zachary Huemann, Graduate Student Researcher, University of Wisconsin-Madison; Xin Tie; Junjie Hu, PhD; Tyler J. Bradshaw, PhD

Introduction
Radiology narrative reports often describe the location, size, and shape of positive findings, but using this descriptive text to guide medical image analysis has been understudied. In this work, we develop a vision-language model for the task of pneumothorax segmentation. Our model, ConTEXTual Net, detects and segments pneumothorax in chest radiographs guided by free-form radiology reports.

Hypothesis
We hypothesized that free-text reports can guide image analysis and improve the performance of artificial intelligence algorithms in segmenting pneumothorax.

Methods
We developed ConTEXTual Net, a multimodal vision-language model which takes in chest radiographs and physician-generated radiology reports and returns pixel-level segmentations of pneumothorax. Vision features are extracted from the image via a U-Net encoding scheme, and language features are extracted via the pretrained language model encoder T5. Cross-attention modules are used to combine the output of each U-Net encoder layer and the text embeddings generated by the pre-trained language model. ConTEXTual Net was trained using the CANDID-PTX dataset consisting of 3,196 positive cases of pneumothorax with segmentation annotations from 6 different physicians as well as clinical radiology reports. We compared ConTEXTual Net to other vision-only models, including a U-Net with a ResNet-50 encoder. We also compared ConTEXTual Net to other vision-language segmentation models, including GLoRIA and LAVT. We performed ablation studies to determine the added value of language integration. Models were compared based on Dice coefficients using five-fold Monte Carlo cross-validation.

Results
ConTEXTual Net (Dice=0.716±0.016) significantly outperformed the vision-only U-Net (0.677±0.015). It also outperformed the vision-language models GLoRIA (0.686±0.014) and LAVT (0.706±0.009). ConTEXTual Net's performance was comparable to the degree of inter-physician variability (0.712 ±0.044) that was computed on a subset of the data. When empty strings were used as the language input to ConTEXTual Net instead of the report, the model's performance (0.671±0.019) dropped significantly. Additionally, we found that altering certain descriptive words regarding location, size, and shape significantly degraded the model performance, demonstrating the strong dependence on language inputs.

Conclusion
Language from free-text radiology reports can be integrated into image analysis models for improved medical image segmentation.

Statement of Impact
This work facilitates the integration of physician input into deep learning-based medical image analysis, enabling better human-in-the-loop artificial intelligence for patient care.

Figure 1. ConTEXTual Net combines a U-Net and transformer architecture. It uses the encoder layers (green) of the U-Net to extract visual representations and uses a pretrained language model (blue) to extract language

representations, then performs cross-attention between the modalities (yellow), and finally uses the decoder layers (red) of the U-Net to predict the segmentation masks.

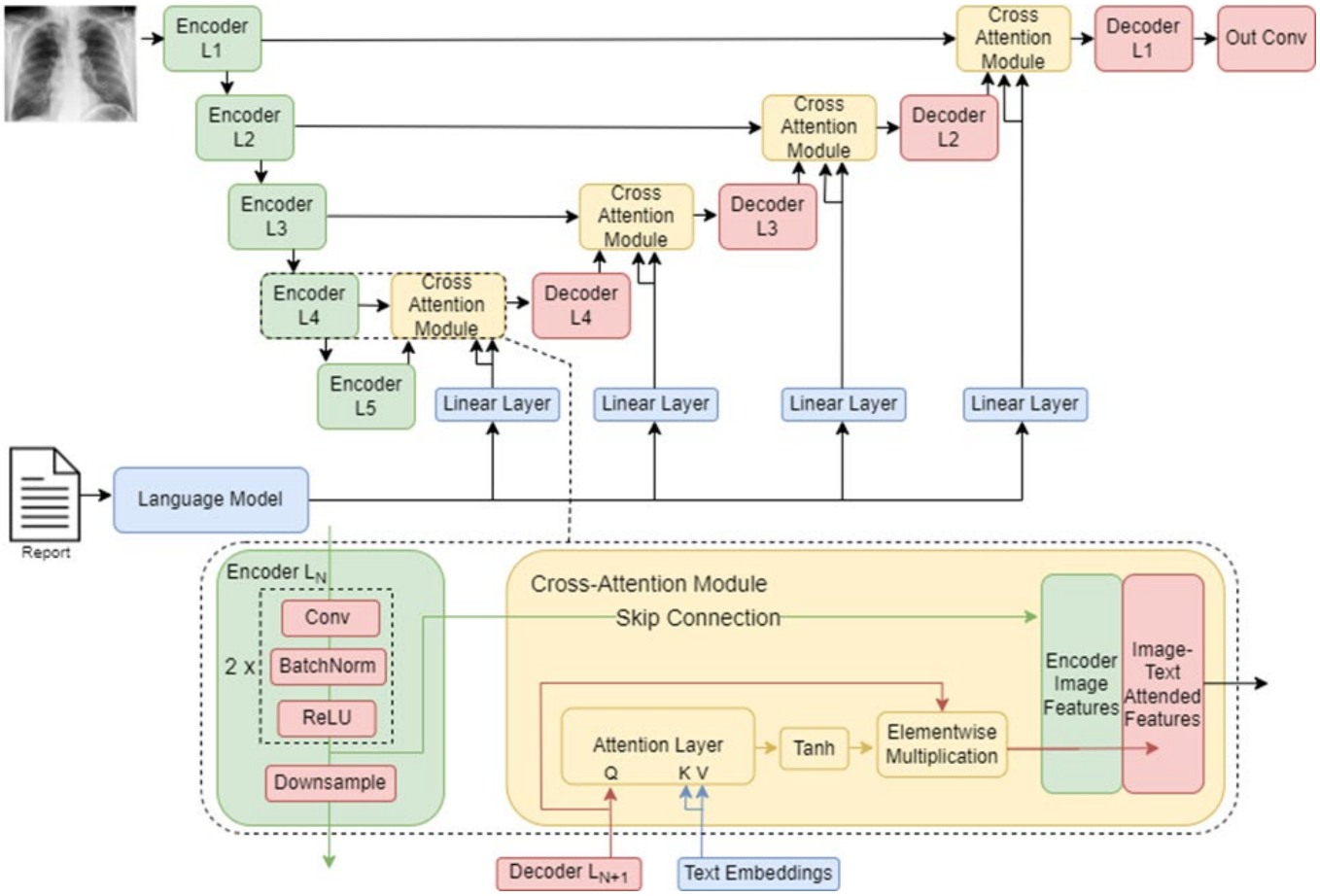


Figure 2. The same input image with different text is fed into the multimodal model. In the top row, an incorrect report describing an apical pneumothorax is used as input with an image, demonstrating that location descriptors like “apical” and “base” carry relevant information for segmentation. In the middle row, we show an example of an image and text with the term “right” changed to “left”. This illustrates the model's sensitivity at the word level. In the bottom row, we changed the term “large” to “small”, which results in a reduction of segmented pixels by 10%. Note “left” and “right” correspond to the patient's “left” and “right”.

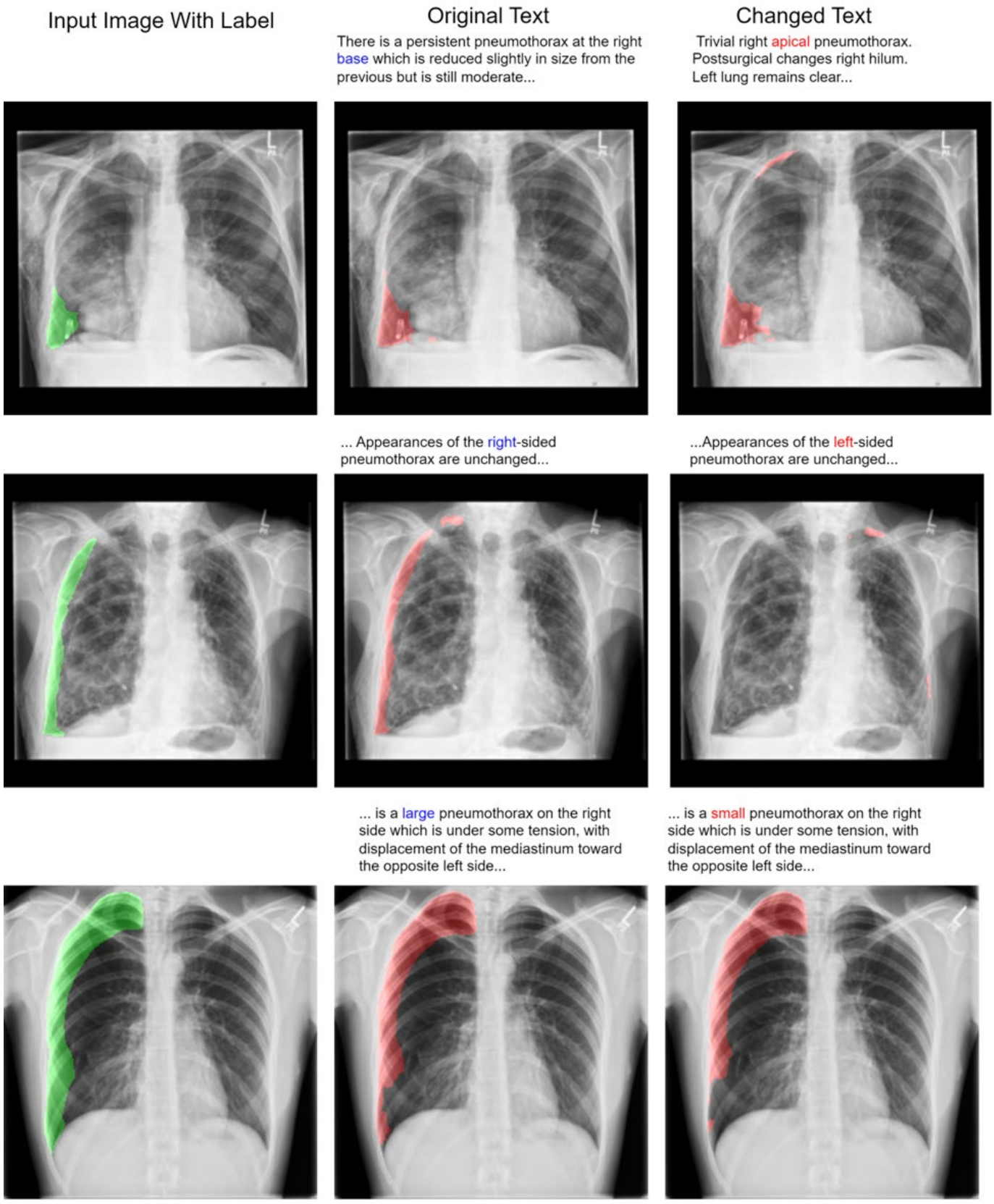


Table 1. ConTEXTual Net was compared to other models and evaluated using ablation studies. Overall, ConTEXTual Net outperformed all other models and variants. Removing the language component of the model resulted in worse performance. Image augmentations were found to improve its performance except for random flipping, which broke the image-text correspondence. We also evaluated the impact of text augmentations, using

different pre-trained language models, using different activation functions within the cross-attention modules, and injecting the text information at different levels of the U-Net.

Model Type	AVG Dice	SD
Model Comparison		
ConTEXTual vision only U-Net	0.687	0.014
Resnet50 U-Net	0.677	0.015
GLoRIA	0.686	0.014
LAVT	0.706	0.009
ConTEXTual Net	**0.716**	0.016
Primary Physician Annotator	0.712	0.044
Vision Augmentations Ablation		
Baseline U-Net w/o augmentations	0.649	0.014
ConTEXTual Net w/o augmentations	0.668	0.010
Baseline U-Net with augmentations	0.687	0.014
ConTEXTual Net with augmentations	**0.716**	0.016
ConTEXTual Net with flipping	0.675	0.016
ConTEXTual Net w/o reports	0.671	0.019
Text Augmentations Ablation		
w/o text augmentations	**0.716**	0.016
Synonym Replacement	0.705	0.008
Sentence Shuffle	0.713	0.023
Synonym + Sentence Shuffle	0.714	0.014
Language Models Ablation		
ConTEXTual Net (T5)	**0.716**	0.016
ConTEXTual Net (Roberta)	0.713	0.010
Activation Functions Ablation		
ConTEXTual Net (Tanh)	**0.716**	0.016
ConTEXTual Net (ReLU)	0.698	0.027
ConTEXTual Net (Sigmoid)	0.710	0.010
ConTEXTual Net (No Activation)	0.704	0.011
Language Injection Ablation		
Attention Module L4	0.712	0.019
Attention Module L3	0.709	0.013
Attention Module L2	0.685	0.021
Attention Module L1	0.679	0.011

Keywords
Multimodal; Vision-Language Models; Pneumothorax; Segmentation

The Brain Tumor Segmentation (BraTS-METS) Challenge 2023

Ahmed W. Moawad, MD, Radiology Resident, Mercy Catholic Medical Center; Anastasia Janas, MD, PhD; Nourel Tahoon, MD; Spyridon Bakas, PhD; Mariam S. Aboian, MD, PhD

Introduction
Brain metastases are the most common malignancy affecting the CNS in adults. The evaluation of brain metastases is commonly limited to comparison to only one prior imaging study. There is a critical need for multi-lesion segmentation and treatment follow-up over multiple studies. Achieving this goal effectively requires the use of automatic algorithms that detect and segment metastases on multiple imaging time points. Accurate detection of small metastatic lesions is critical for patient prognosis. Many of the segmentation algorithms that were developed for gliomas demonstrate high accuracy for larger metastases, but their performance for small metastases is lower. Addressing this challenge is critically important for the development of novel segmentation and detection algorithms specifically designed for brain metastases that are common in clinical practice.

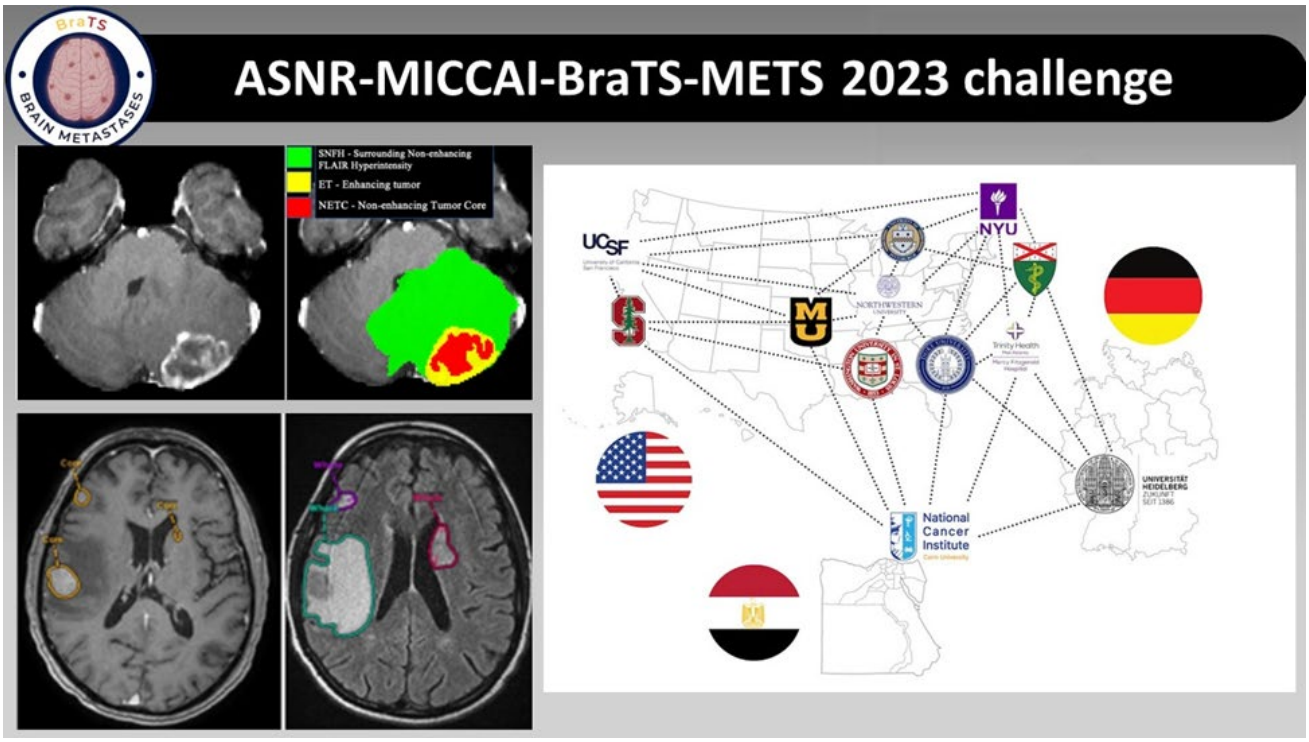
Hypothesis
NA

Methods
The BraTS-METS 2023 dataset is acquired from patients scanned on varying MRI imaging quality across different vendors. The scans are pre-processed using different algorithms then refined by a pool of annotators. The dataset is finally reviewed carefully by two independent board certified neuroradiologists. The data is divided into Training, validation and testing dataset. BraTS-METS 2023 challenge will be evaluated based on multiple parameters.

Results
There are 12 different institutions over three continents participated in the BraTS-METS 2023 dataset. We received > 2500 multi-parametric MRIs which underwent preprocessing to wipe out all PHI, reorient and be consistent with all BraTS space and header. Initial raw data were pre-segmented with three different algorithms then fused together to get consensus pre-segmentation file. A pool of 150 annotators and 50 board certified attendings were recruited through ASNR mass call for volunteer announcement. Studies are first assigned to annotators, reviewed by 1-2 board certified neuroradiologists then reviewed by single senior neuroradiologist for consistency and quality control final check. Training and Validation datasets are made available to public through BraTS-METS 2023 website. The whole project is part of TCIA/NCI moonshot program.

Conclusion
Development of accurate segmentation algorithms to detect small metastasis is crucial for radiation planning and has the potential to improve care of patients with brain metastasis. MICCAI ASNR BraTS-METS Challenge is an important initiative to include multi-institutional international datasets in order to develop general model applied to all patients with brain metastasis from any institution. Our study has several limitations include establishing culture of sharing among various institutions, Tedious process of manual annotations among students with relatively limited knowledge of differ imaging features of brain metastasis.

Statement of Impact
NA



Left Panel: This is an example of tumor sub-regions annotated in the different mpMRI scans. right panels shows map of institutions that reached out with interest in contributing data to the ASNR MICCAI BraTS-METS Brain Metastases Challenge.

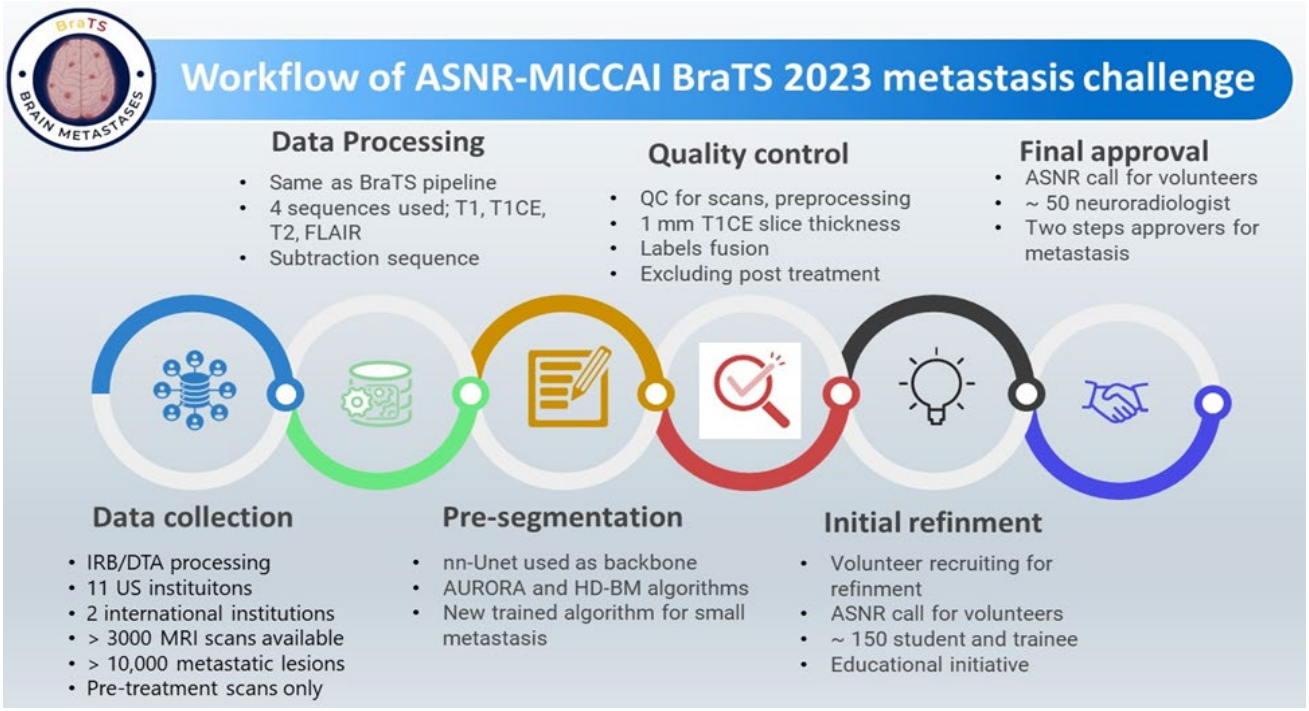


Figure shows the workflow used in BraTS-Met 2023 challenge.

Keywords
BraTS-MET; Openscience; Segmentation; Challenge

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Using An Open-source Language Model to Abstract the Presence of Acute Cervical Spine Fracture from Radiologic Reports: A HIPAA Compliant Alternative to "ChatGPT"

Bardia Khosravi, MD, MPH, MHPE, Postdoctoral Research Fellow, Mayo Clinic AI Lab; Sanaz Vahdati, MD; Pouria Rouzrokh, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Ali Ganjizadeh, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction

ChatGPT is a large language model (LLM) with a chat interface that allows for back-and-forth communication with users. While the engineering behind aligning the model with human values is a significant advantage, the closed-source nature of the model and the need to transmit sensitive patient information through a commercial API make it unsuitable for medical data abstraction. Additionally, all studies that have used ChatGPT to extract radiologic reports have created synthetic reports to comply with HIPAA regulations. As a result, open-source LLMs may be a more viable option for data abstraction from radiologic reports because they can be deployed locally.

Hypothesis

In this study, we use an open-source LLM to determine the presence and level of cervical spine fractures in radiologic reports.

Methods

We collected radiologic reports from 115 non-contrast cervical spine CT scans taken between January and February 2022. Of these reports, 15 were used for prompt engineering and fine-tuning, while the remaining 100 were used to test the model's performance. The model's objective was to generate a JSON file indicating the presence or absence of an acute cervical vertebral fracture and its associated level. This particular problem was chosen because the model must distinguish between different anatomical regions and differentiate between chronic and acute fractures. We used an open-source LLM, specifically the Large Language Model Meta AI (LLaMA-13B variation), and deployed it locally. We employed a combination of chain-of-thought (CoT) and few-shot prompting techniques, asking the model to produce a final JSON output of the results (Figure 1).

Results

The model consistently produced correctly formatted JSON outputs. It achieved a 99% (99/100) accuracy rate in detecting acute cervical spine fractures. Notably, in 16 out of 17 cases where fractures were present, the model accurately predicted the level of the fracture. However, in five instances where multiple vertebrae levels were fractured, the model only predicted one.

Conclusion

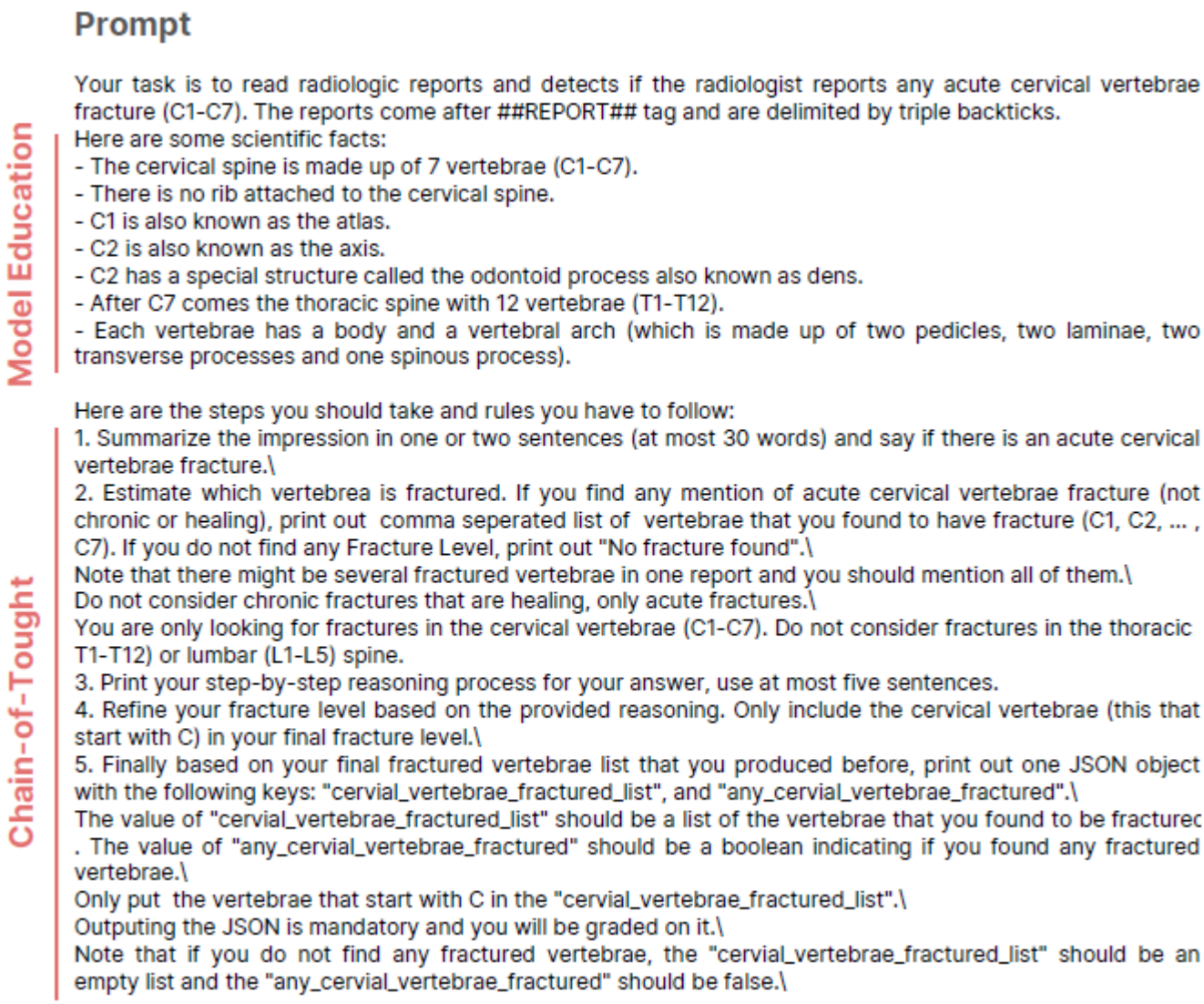
With appropriate prompting, open-source LLMs can perform well without the need to share patient data with

external sources.

Statement of Impact

LLMs can be used to extract information from unstructured text. While ChatGPT has gained significant popularity, it only offers a commercially hosted API, which requires transmitting patient information outside of the healthcare institution. It is important to explore open-source alternatives and their capabilities in sensitive scenarios involving protected health information (PHI).

Figure 1. The final prompt to the model.



Output format constraint

here is how you should format your output:\
<Summary>
* your summary *
</Summary>
<Initial Fracture Level>
* list of Fracture Level or if none type No fracture found *
</Initial Fracture Level>
<Reasoning>
* your reasoning *
</Reasoning>
<Final Fracture Level>
* your final Fracture Level *
</Final Fracture Level>
<JSON>
* your JSON output *
</JSON>

Few-shot example 1

~~~
## REPORT ##
```1. Acute type II odontoid fracture with posterior displacement of the odontoid and associated prevertebral hematoma. No significant spinal canal narrowing. 2. Multilevel, moderate spondylotic changes of the cervical spine. ```

<Summary>
There is an acute type II odontoid fracture.
</Summary>
<Initial Fracture Level>
C2
</Initial Fracture Level>
<Reasoning>
Based on the report, there is an acute type II odontoid fracture in the report. The odontoid is part of C2. So C2 has an acute fracture.
</Reasoning>
<Final Fracture Level>
C2
</Final Fracture Level>
<JSON>
{
  "any_cervial_vertebrae_fractured": true,
  "cervial_vertebrae_fractured_list": ["C2"]
}
</JSON>
~~~

Few-shot example 2

~~~
## REPORT ##
```1. Abnormal widening of the right C2-C3 facet joint without definite associated fracture. Findings are concerning for possible underlying associated ligamentous injury. 2. Advanced spondylotic changes throughout the cervical spine. ```

<Summary>
There is no mention of acute cervical vertebrea fracture.
</Summary>
<Initial Fracture Level>
No fracture found
</Initial Fracture Level>
<Reasoning>
Based on the report, there is a mention of abnormal widening of the right C2-C3 facet joint without definite associated fracture, but there is no acutely fractured cervical vertebrae.
</Reasoning>
<Final Fracture Level>
No fracture found
</Final Fracture Level>
<JSON>
{
  "any_cervial_vertebrae_fractured": false,
  "cervial_vertebrae_fractured_list": []
}
</JSON>

Question

~~~
REPORT
```%%REPORT%%```

<Summary>

Keywords
Natural Language Processing; Large Language Models; Privacy; Open Science
```

Using Graph Representation Learning to Passively Learn Imaging Protocols

Dimitri Falco, PhD, AI Engineer, Quantivly; Robert MacDougall, MS; Benoit Scherrer, PhD

Introduction

Imaging protocol management is a complex, unsolved problem. Systems like HL7 or DICOM only record the "general procedure type" (e.g., "MR BRAIN W/WO") but not the specific protocol performed (e.g., Stroke, TBI, Pediatric Tumor), hindering data-driven monitoring and optimization. Worse: protocols, the exams' "recipe", can in reality be explicit (e.g. in a book) or implicit (remembered or modified by a technologist); and the information location is ill-defined. We propose to use machine learning to passively learn imaging protocols from the technical parameters in DICOM metadata, leading to a more accurate description and new protocol management capabilities.

Hypothesis

We previously built a software platform that analyzes DICOM metadata on-the-fly to construct a new, highly granular ontology of imaging operations. Key features include the extraction of new concepts beyond DICOM including the concepts of examinations, acquisitions and volumes, and access to all technical parameters (e.g., TE, TR, resolution, etc) of each volume. Our hypothesis is that we can use this data platform to automatically learn imaging protocols with machine learning by analyzing the technical parameters of acquisitions.

Methods

Passively learning imaging protocol amounts to automatically grouping together exams with similar acquisitions. This is a difficult unsupervised clustering problem as different exams have different numbers of acquisitions - requiring computing distances between spaces of different dimensions. We propose, for the first time, to represent exams as graphs in which each node is an acquisition with its technical parameters as node attributes. We then use graph neural networks to automatically learn an embedding for each exam that can be used for many downstream machine learning applications.

Results

After training our graph neural network and using uniform manifold approximation and projection (UMAP) to visualize our embeddings, we found that our embedding could successfully, geographically regroup similar exams (Figure 1). Importantly, our model produced similar embeddings in presence of repeated acquisitions within examinations, but different embeddings when acquisitions were missing (Fig 2). Figure 3 shows that community analysis in regions of the manifold can further divide the data into distinct subprotocol communities, solely based on technical parameters.

Conclusion

We show that transforming imaging exam metadata into graphs and subsequently training a graph neural network, allows us to passively learn imaging protocols simply from the technical parameters of the acquisitions.

Statement of Impact

Passive protocol learning will radically transform imaging operation management allowing new deviation detection capabilities, protocol normalization across scanners, and the assessment of protocol-specific statistics crucial for optimizing patient scheduling.

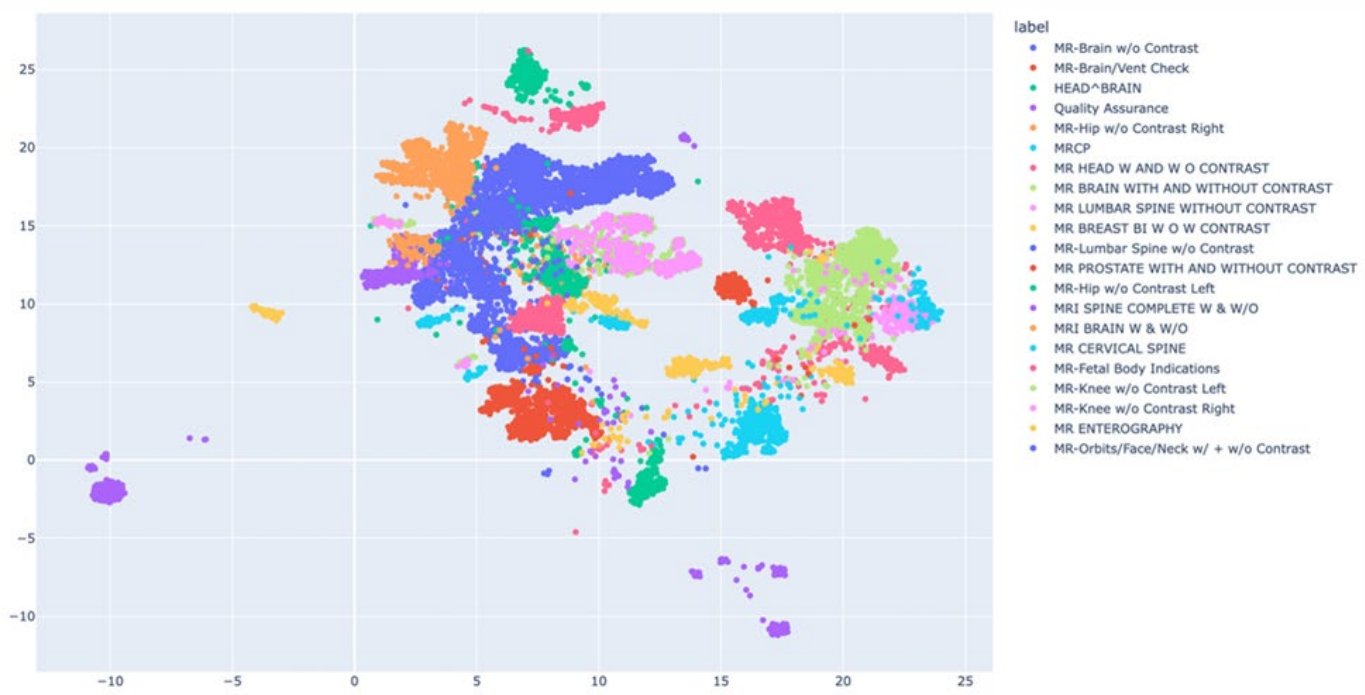


Figure 1: UMAP dimension reduction of GNN exam embedding. Here we show that different anatomical areas identified by their “Study Description” are in distinct locations, and that exams of the same study description are spread out in specific patterns representing distinct sub-protocols. Importantly we show that lateral protocol (left vs right) occupies the same space and that unique exams (like Quality Assurance) as in distinct areas. This shows our GNN is able to translate imaging exam metadata into distinct embeddings that hold semantic information unique to each protocol.

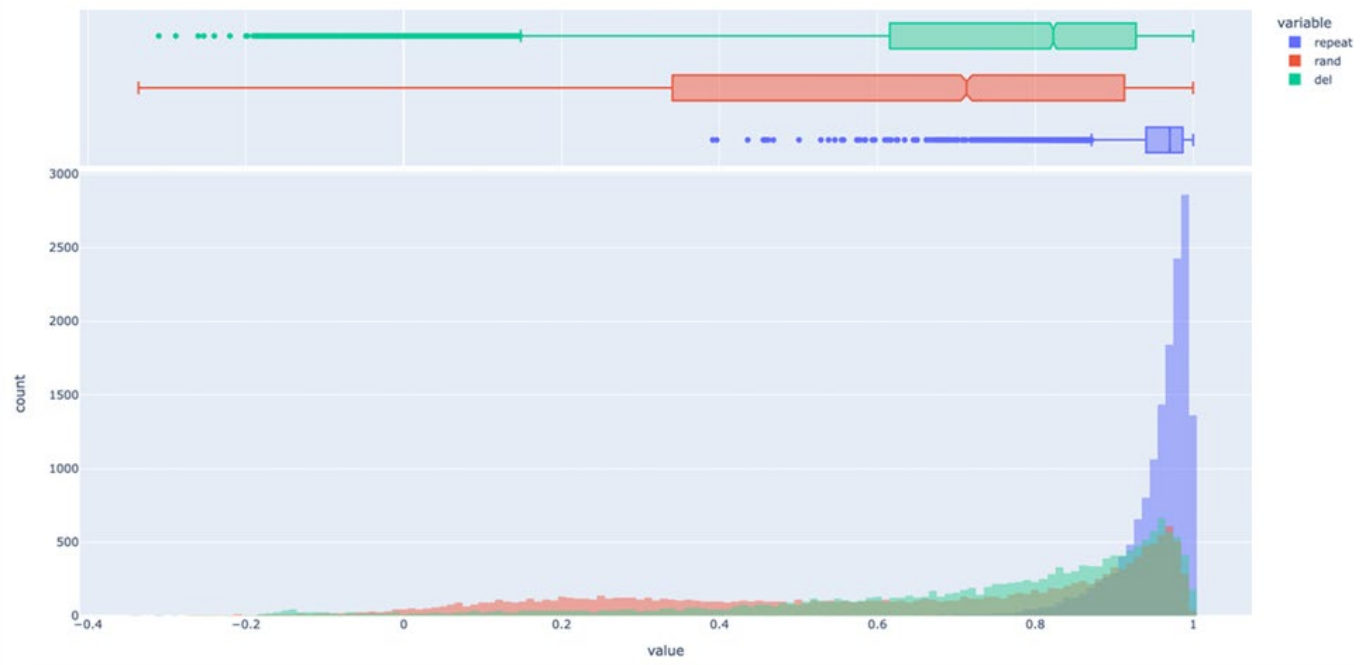


Figure 2: In order to determine if similar embeddings are produced, we computed the correlation between different pairs of embeddings: 1) normal embeddings (unaltered graph) and artificial repeat; 2) normal embeddings and deleted acquisition; 3) normal embeddings and augmented graph (where the added node have random attributes). The distributions of correlation values show a high level of similarity in presence of repeated acquisitions, indicating robustness of our embedding to repeats. As expected, the correlations are much lower with other perturbations which is promising to identify protocol deviations.

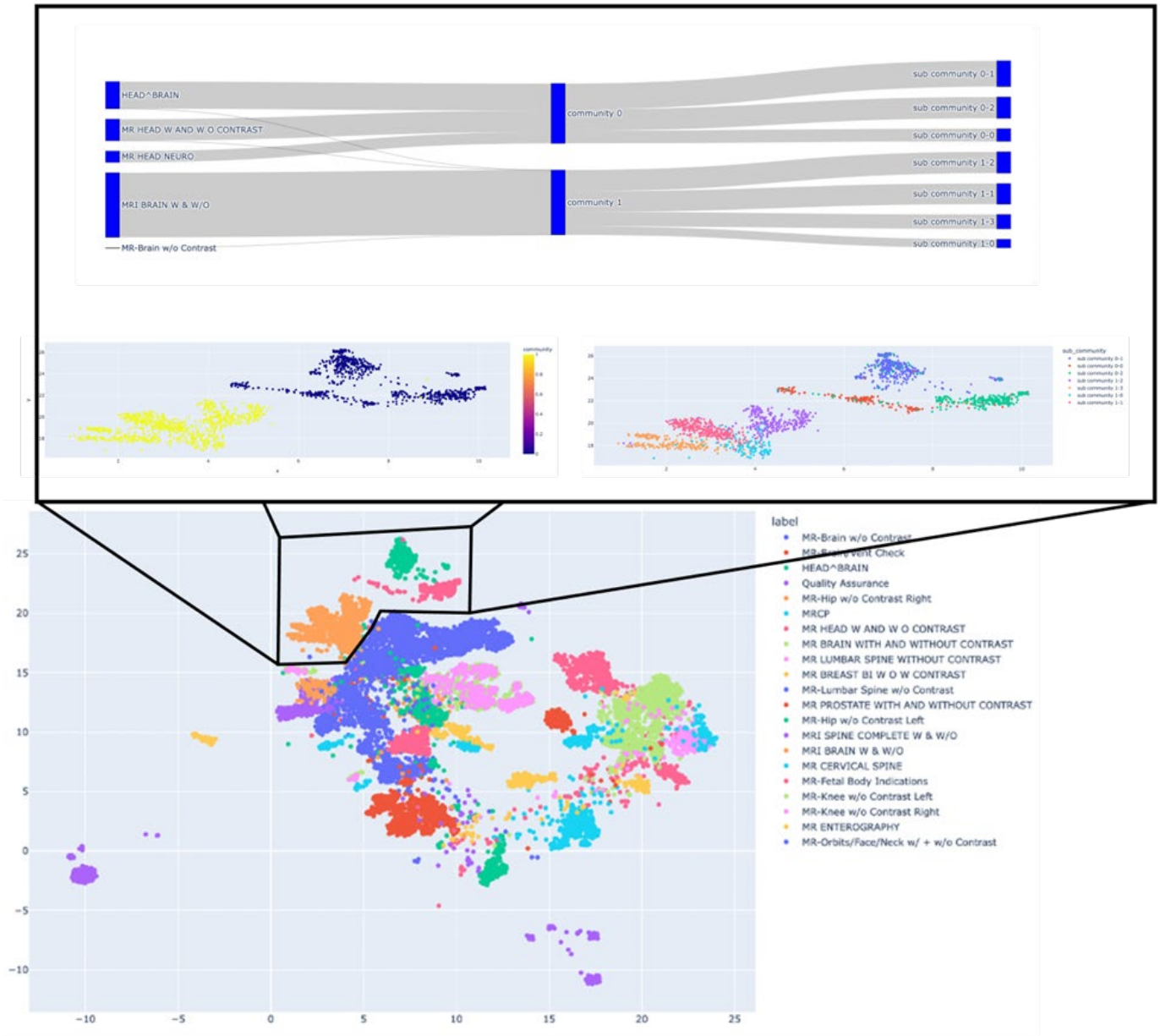


Figure 3: Community analysis (top part) of all exams found in the highlighted region (bottom part). The community analysis was based on a combination of the Louvain Community Detection Algorithm and a modified version of the Jonker-Volgenant algorithm. The results indicate that our model has the ability to isolate clusters and sub-clusters that represent distinct imaging protocols based solely on technical parameters, effectively demonstrating protocol learning

Keywords
Imaging Protocol learning; Graph Neural Network; Graph Representation Learning

WESTERN-RLP: Augmenting Image-Caption Radiology Datasets using Image Embeddings Search of Large-Scale Natural Image Databases

Kartik Gupta, Medical Student, University of Western Ontario; David Li, MD; Jaron Chong, MD

Introduction

Recent breakthroughs in multimodal Large Language Models (LLM's) potentiate advanced Visual Question Answering (VQA) models in clinical radiology, requiring however, numerous high-quality text captioned images to train. Open-source datasets such as Radiology Objects in COntext (ROCO), are modest in size (N=65,419), dwarfed by much larger natural image datasets derived from large Internet web crawls, such as LAION-5B (N=5.85 Billion). We investigate the potential dataset augmentation yield of searching large natural image databases for radiology relevant images, utilizing vector embedding image similarity searches.

Hypothesis

Vector embedding database similarity searches of natural image databases can augment existing radiology image datasets.

Methods

65,419 images from the ROCO dataset were input to a pre-computed KNN ViT/L-14 embeddings search of LAION-5B. From every ROCO image, the Top 100, 1000, and 10000 result depth vector searches from LAION-5B were processed and deduplicated (i.e. D100, D1000, D10000). To pre-categorize the image sets, an image classifier (MobileNet-v2) was trained on 3500 “Relevant”, 3426 “Conditional Relevant (Usable With Post-Processing)”, and 2748 “Non-Relevant” images sampled from ROCO & Non-Relevant LAION-5B images. The augmented image sets were randomly sampled (5 samples; 250 images), to perform manual category validation and confirm automated estimates.

Results

Augmented similar image searches resulted in 133,960, 515,220, and 1,180,148 unique images for D100, D1000, and D10000 searches respectively. The MobileNet-v2 image classifier in a held-out ROCO/LAION-5B test group had a 93% accuracy and 79% F1 score. Automated estimated absolute and percent yields over ROCO inputs for combined Relevant and Conditionally Relevant classes for the D100, D1000, and D10000 searches were 57,954 (+89%), 122,276 (+187%), and 185,021 (+283%) images respectively. This was concordant with scaled yields derived from manual sub-sampling validation which estimates a combined Relevant and Conditional Relevant count of 55,727-71,802 (+85-110%), 86,557- 142,200 (+132-217%), and 103,853-207,706 (+159-318%) additional images respectively. Sources of non-relevant contamination images included: anatomic illustration, medical artistic renderings, surgical images, veterinary radiographs, and electronic microscopy images.

Conclusion

In this study, we demonstrate a novel form of image dataset augmentation, by utilizing vector embedding searches of natural image datasets. Our estimates show up to a 1.5-3 times augmentation on the ROCO dataset. This method is potentially generalizable to any pre-existing set of curated radiology images.

Statement of Impact

This study demonstrates the feasibility of using vector embedding searches on natural image datasets to augment existing medical image datasets.

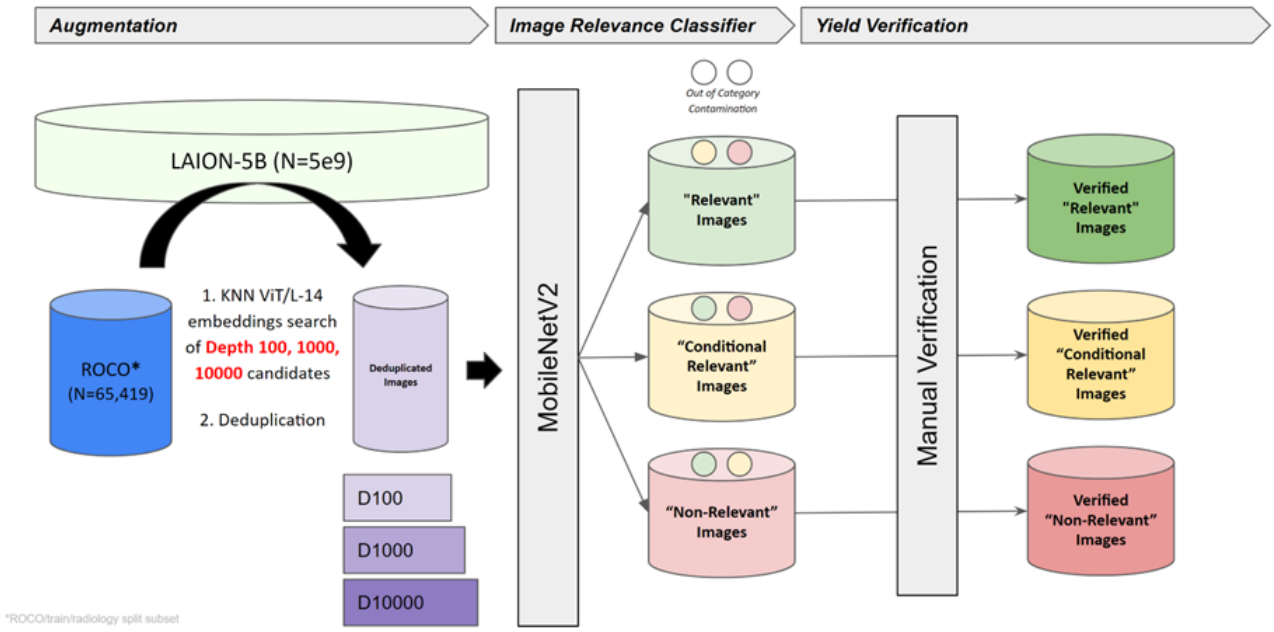


FIGURE 1: ROCO to LAION-5B Augmentation Pipeline and Yield Estimation procedure. Utilizing pre-computed ViT/L-14 embeddings of the LAION-5B database, image similarity searches were performed using individual ROCO radiology images into the greater LAION-5B dataset, at various search result depths of Top 100, 1000, and 10000 images, which were duplicated to generate 3 new augmented datasets of D100, D1000, and D10000, with sizes ranging from 133K to 1.1M candidate images. These images were subjected to further yield estimate analysis via both automated (MobileNetV2) and manual random sample verification.

		Manually Verified Random Sample Classifications					
		Random Sample 1 (N=250)	Random Sample 2 (N=250)	Random Sample 3 (N=250)	Random Sample 4 (N=250)	Random Sample 5 (N=250)	Average (Std Dev)
Relevant	D100	75	59	65	66	61	65 (6)
	D1000	22	23	27	35	18	25 (6)
	D10000	22	11	14	11	10	14 (5)
Conditional Relevant	D100	45	50	59	57	52	53 (5)
	D1000	33	34	32	24	34	31 (4)
	D10000	13	22	12	18	17	16 (4)
Non-Relevant	D100	133	141	126	127	137	132 (6)
	D1000	195	193	191	191	198	193 (3)
	D10000	215	217	224	221	223	220 (4)

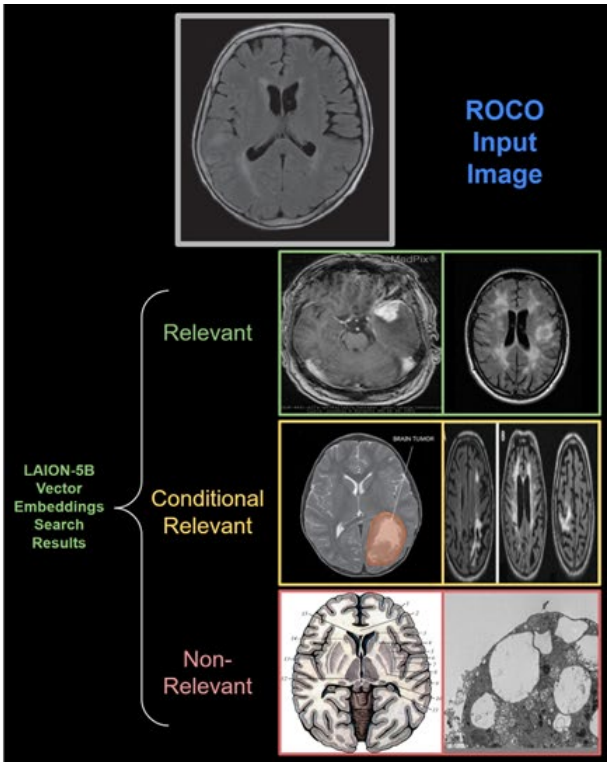
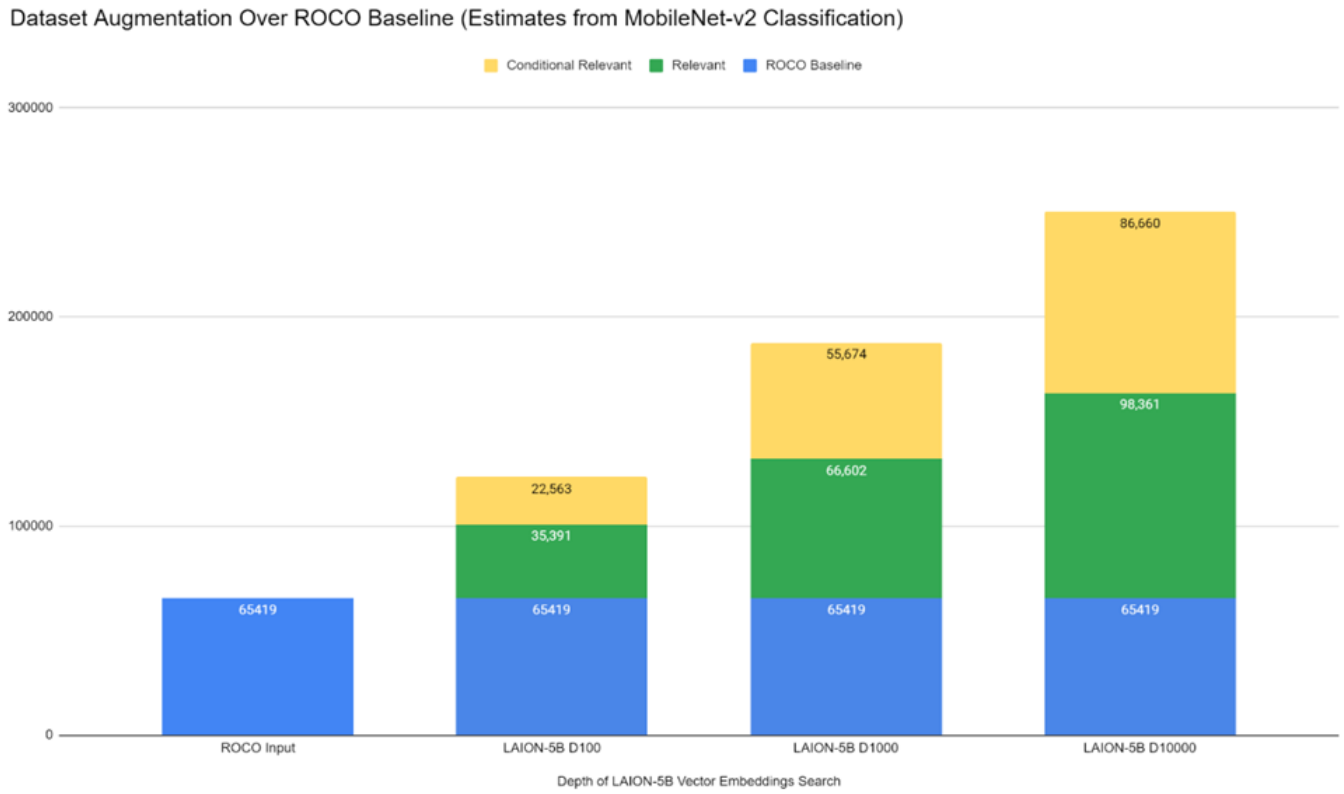


FIGURE 2: Five random samples of 250 images were taken from D100, D1000, and D10000, and were subjected to manual review. Pertinent criterion for determining relevance, particularly with an aim of further training a Multi-

Modal LLM or VQA model, was assessed by evaluating for image quality, obscuration of findings by annotations, radiology subject matter appropriateness, and whether an image was single versus part of a multi-panel set of images. Many conditional relevant images, were deemed conditional on the basis of being part of a multi-panel set, and were felt to be utilizable subject to additional post-processing and cropping. Non-relevant images commonly included artistic or anatomic renderings, and images with similar visual characteristics such as scanning electron microscopy or veterinary radiology images.



Augmentation Relevant and Conditional Relevant LAION-5B Images over Baseline ROCO Input

LAION-5B Search Depth	Relevant	Conditional Relevant	Total
ROCO Input	65,419 (100.0%)	---	65,419
LAION-5B D100	35,391* (26.4%)	22,563* (16.8%)	57,954*
LAION-5B D1000	66,602* (12.9%)	55,674* (10.8%)	122,276*
LAION-5B D10000	98,361* (8.3%)	86,660* (7.3%)	185,021*

**Categorical labels estimated from Automated MobileNet-v2 Relevance Classifier*

FIGURE 3: Relationship between increasing search depth and additional augmentation over ROCO input as estimated by the automated MobileNet-v2 relevance classifier. Although the percentage yield of Relevant and Conditional Relevant classes showed step-wise decreases with increasing search depth, the absolute number of relevant images continued to increase. At a search depth of 10,000, the potential augmentation yield is estimated to be 1.5 (Relevant Only) to 3X (Relevant+Conditional Relevant) that of the original ROCO input.

Keywords
Multi-modal Large Language Model; Image Captioning; Visual Question Answering

Networking Reception & Scientific Abstract Poster Discussions

Date:
SUN, OCT 1

Time:
6:30 PM – 7:45 PM ET

Location:
Turner Hall & Pre-function Area

Posters are displayed in the order of this listing.

Adaptive Deep Learning for Precise Early Stage Lung Tumor Delineation on 4D imaging
+ Luis Ricardo de la O Arevalo, MS, PhD Student, University Medical Center Groningen
+ Nanna M. Sijtsema, PhD; Alessia De Biase, MS; Johannes A. Langendijk, PhD; Robin Wijsman, PhD; Peter M.A. van Ooijen, PhD

Advancing Hepatic Decompensation Status Prediction through Computed Tomography-Based Radiomics Signature and Body Composition Model Integration
+ Yashbir Singh, PhD, Assistant Professor, Mayo Clinic
+ John Eaton, MD; Sudhakar Venkatesh, MD; Bradley Erickson, MD, PhD, CIIP, FSIIM

An Interactive Decision Support Tool for Evaluating Machine Learning Algorithm Performance in Medical Image Analysis developed by the Medical Imaging and Data Resource Center (MIDRC): MIDRC-MetricTree
+ Tingting Hu, PhD Candidate, Visiting Scientist, U.S. Food & Drug Administration
+ Natalie Baughan, PhD; Karen Drukker, PhD, MBA; Maryellen L. Giger, PhD; Grace Hyun Kim, PhD, MS; Mike McNitt-Gray, PhD; Berkman Sahiner, PhD; Emily Townley, Heather Whitney, PhD

Analysis of Intersectional Bias in a Novel Melanoma Image Classification Algorithm
+ Christopher Caligiuri, Student, Princeton University

ChatGPT Enhanced Radiology Reporting using PRECISE Framework For Patient-Centered Care
+ Satvik Tripathi, Research Student, Massachusetts General Hospital
+ Emiliano Garza, MD; Liam Mutter; Michael Dezube, MEng; Christopher P. Bridge, PhD; Dania Daye, MD, PhD

Deep Learning-Based Natural Language Processing for Classification of Renal Surgical Pathology Outcomes in a Multi-Site Dataset
+ Satvik Tripathi, Undergraduate Researcher, Perelman School of Medicine at the University of Pennsylvania
+ Rithvik Sukumaran; Jackson Steinkamp; Charles M. Chambers; Darco Lalevic; Hanna M. Zafar, MD, MHS; Tessa S. Cook, MD, PhD, CIIP, FSIIM, FCPP

Deep Learning Assisted Curation of the CANDID-III Dataset with Free-text Reports
+ Anna Hu, Medical Student, George Washington University School of Medicine and Health Sciences
+ Sijing Feng, MBChB; Qixiu Liu, MBChB; Darren Ritchie, MBChB; Benjamin Wilson, MBChB, FRANZC

Evaluation of ChatGPT Performance on Radiology Board Exam-Style Questions
+ Anna Hu, Medical Student, George Washington University School of Medicine and Health Sciences
+ Sijing Feng, MBChB; John Egbuji, MBChB; Sophia Dean, MBChB; Ben K. Wilson, MBChB

Decoder-Only Computed Tomography Radiology Reports (DOCTRR)
+ Tegan Keigher, MS, Graduate Student, Data Science Institute, University of Chicago
+ Benan Akca, PhD; Andrew Alvarez, MS; Mary Erikson, MS; Utku Pamuksuz, PhD

Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension
+ Jen Park, MS, PhD Student, Johns Hopkins University
+ Kyulee Jeon; Haridimos Kondylakis, PhD; Teri Sippel Schmidt, MS, FSIIM; Paul Nagy, PhD, CIIP, FSIIM; Seng Chan You, MD, PhD

Do General Purpose Large Language Models Outperform Domain-Specific NLP Methods for Radiology Report Label Extraction?
+ Cody Savage, MD, Radiology Resident, University of Maryland Medical Intelligent Imaging (UM2ii) Center
+ Steven Rothenberg, MD; Andrew D. Smith, MD, PhD; Paul H. Yi, MD, MS

Enhancing Efficiency and Performance in Healthcare: A Federated Learning Approach for CT Image Segmentation
+ Alan McMillan, PhD, Professor of Clinical Health Sciences, UW Health
+ Iman Z. Estakhraji, PhD; John W. Garrett, PhD; Kristopher Kersten

Enhancing Radiology Reports and Medical Information Retrieval through Question and Answering with Large Language Models
+ Nitin Gupta, MS, Student, University of Chicago
+ Ananth Prayaga, MS; Jialin Wu, MS; Hsin-Yen Yeh, MS; Utku Pamuksuz, PhD; Benan Akca, PhD

Evaluating the Diagnostic Performance of a Deep Learning Model for Detecting Thyroid Nodule Malignancy: An Expert Evaluation Study
+ Sanaz Vahdati, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Bardia Khosravi, MD, MPH, MHPE; Kathryn Robinson, MD; Pouria Rouzrokh, MD, MPH, MHPE; Mana Moassefi, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Extraction of Labels from Radiology Reports using ChatGPT
+ Jason Adleberg, MD, Radiology Resident, Mount Sinai
+ Nicholas Primiano, MD; Alexander Kagen, MD; Tessa S. Cook, MD, PhD, CIIP, FSIIM

From the Operating Room to the Office: Digital Staining of White Light Cystoscopy Videos for Bladder Cancer Detection
+ Greyson Wintergerst, Undergraduate Researcher, Vanderbilt University
+ Shuang Chang, PhD; Haoli Yin; Kristen R. Scarpato, MD, MPH; Amy N. Luckenbaugh, MD; Sam S. Chang, MD, MBA; Soheil Kolouri, PhD; Audrey K. Bowden, PhD

Generation of Radiology Report's Impression from Findings' Description on Pediatric Abdomen Ultrasound
+ Dana Alkhulaifat, MD, Postdoctoral Research Fellow; Children's Hospital of Philadelphia
+ Vahid Khalkhali, PhD; Patricia P. Rafful, MD, PhD; Michael Welsh, PhD; Susan T. Sotardi, MD

Large Language Model Improves Detection of Negated Expressions in Radiology Reports
+ Yonatan Babore, Medical Student, University of Pennsylvania
+ Yvonne Su; Charles E. Kahn, MD, MS, FACR, FSIIM

Identifying Cerebrospinal Fluid Leak using Brain MRI: A Deep Learning Approach
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Mana Moassefi, MD; Ian T. Mark, MD; Ajay Madhavan, MD; Jared T. Verdoorn, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; John C. Benson, MD

Towards Trustworthy Deep Learning: Applying Mondrian Conformal Prediction to Intracranial Hemorrhage Detection
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Cooper Gamble

Uncertainty Quantification in Radiogenomics Analysis Using Mondrian Conformal Prediction
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Mana Moassefi, MD; Khosravi Bardia, MD, MPH, MHPE; Gian Marco Conte, MD, PhD; Pouria Rouzrokh, MD, MPH, MHPE;; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Leveraging 3D Segmentation Datasets for Rapid Body Region Classification
+ Xue Li, MS, Research Assistant, University of Wisconsin-Madison
+ Nicolas Pannetier, PhD; Mehul Sampat, PhD; Travis Richardson; Richard Bruce, MD; John W. Garrett, PhD; Alan B. McMillan, PhD

mRMR-permute: Permutation Testing for Unbiased Minimum Redundancy Maximum Relevance Feature Selection
+ Winston T. Chu, PhD, Data Scientist, Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health
+ Marcelo A. Castro, PhD; Venkatesh Mani, PhD; Jens H. Kuhn, M.D., PhD; Ian Crozier, MD; Claudia Calcagno, PhD; Jeffrey Solomon, PhD

One Copy Is All You Need: Resource-Efficient Streaming of Medical Imaging Data at Scale
+ Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine
+ Adway Kanhere, MS; Eliot L. Siegel, MD, FSIIM; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Text2Cohort: Democratizing the NCI Imaging Data Commons with Natural Language Cohort Discovery
+ Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine
+ Adway Kanhere, MS; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Pruning and Principal Component Analysis (PCA) on UNet++ for Segmentation of Kidneys and Cysts in Autosomal Dominant Polycystic Kidney Disease (ADPKD)
+ Chetana Krishnan, Graduate Student, University of Alabama at Birmingham
+ Emma Schmidt; Ezinwanne Onuoha, MS; Michal Mrug, MD; Carlos E. Cardenas, PhD; Harrison Kim, PhD

SegViz: A federated Learning Framework to Train Multi-task Segmentation Models from Partially Annotated and Distributed Datasets
+ Adway U. Kanhere, MS, Software Engineer, University of Maryland School of Medicine
+ Pranav S. Kulkarni; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Stimulated Raman Histology Image Reconstruction Using Weakly Supervised Generative Adversarial Networks
+ Sung Jik Cha, Medical Student, Western Michigan University
+ Yiwei Liu, MS; Esteban Urias, MD; Christian Freudiger, PhD; Todd Hollon, MD

Unveiling Segmentation Errors: Enhancing Auto-Segmentation with ML Models Trained on Radiomic Features

+ Abishek Karki, PhD, Research Associate, University of Virginia
+ Victor G. Leandro Alves, PhD; Hashir N. Rashad, PhD; Jeffrey V. Siebers, PhD

Using MONAI Pre-Trained Models for Colorectal Tissue Type Phenotyping: A Feasibility Study to Integrate Deep Learning Model Results using the Medical Extension OMOP CDM

+ Shijia Zhang, PhD Student, Johns Hopkins University
+ WooYeon Park, MS; Blake Dewery, PhD; Paul Nagy, PhD, CIIP, FSIIM

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Adaptive Deep Learning for Precise Early Stage Lung Tumor Delineation on 4D imaging

Luis Ricardo de la O Arevalo, MS, PhD Student, University Medical Center Groningen; Nanna M. Sijtsema, PhD; Alessia De Biase, MS; Johannes A. Langendijk, PhD; Robin Wijsman, PhD; Peter M.A. van Ooijen, PhD

Background/Problem to be solved

Stereotactic Body Radiotherapy (SBRT), is the standard of care for inoperable early stage Lung tumors. However, SBRT is affected by respiratory motion. To ensure accurate treatment, the tumor motion is captured using 4DCT, which contains the tumor position throughout the breathing phases (BP). Current clinical practice is manual delineation of the Gross Tumor Volume (GTV) in a few or all breathing phases of the 4D-CT. Then GTVs of all breathing phases are combined to generate the Internal Target Volume (ITV).

Intervention(s)

The dataset of 222 scans was acquired at the University Medical Center Groningen as routine for radiotherapy treatment planning. Each record consisted in a 4D-CT scan (10 BPs and 1 average scan), a GTV contour in 50% BP and one ITV contour in the average scan. Data was split into training (56%), validation (20%) and test set (24%). Two neural networks, SwinUNetR and DynUNet were trained to delineate the GTV on the 50%BP. Additionally, by merging the two outputs of the networks with a logical OR operation a third output was tested. To create the ITV, the trained network generated GTV contours for the 9 remaining BPs. Two methods were compared to assemble the ITV. First, by creating a polygon containing all 10 GTVs (10BP). Second, using only the max inhalation (0%) and max exhalation (50%) breathing phases (2BP).

Outcome

Performance metrics included the Dice Score (DSC), Hausdorff distance 95th percentile (HD95), among others. For the GTV segmentation, SwinUNETR had better metrics overall, indicating a higher detection rate. However, when considering only true positive detections, DynUNET showed higher similarity metrics. Nevertheless, the combination of both networks yielded the highest performance, achieving a DSC of 0.80±0.14 and HD95 of 4.33±4.08 mm in the test set. Test set ITV segmentation achieved a DSC of 0.65±0.09 and 5.57±2.17 mm HD95 when using 10BP, and a DSC of 0.69±0.09 DSC and 5.04±2.22 mm HD95 in the 2BP. A correlation was observed where the higher GTV DSC value corresponded to larger DSC for the ITV, and smaller HD95 for the GTV corresponded to smaller HD95 for the ITV.

Conclusion

This study demonstrated the ability of the SwinUnetR + DynUNet approach to delineate the GTV in a specific trained breathing phase with a DCS of 0.8, as well as other breathing phases to construct an ITV with a DCS of 0.65-0.69.

Statement of Impact

This method shows promise for automating tumor contouring in early stage lung cancer.

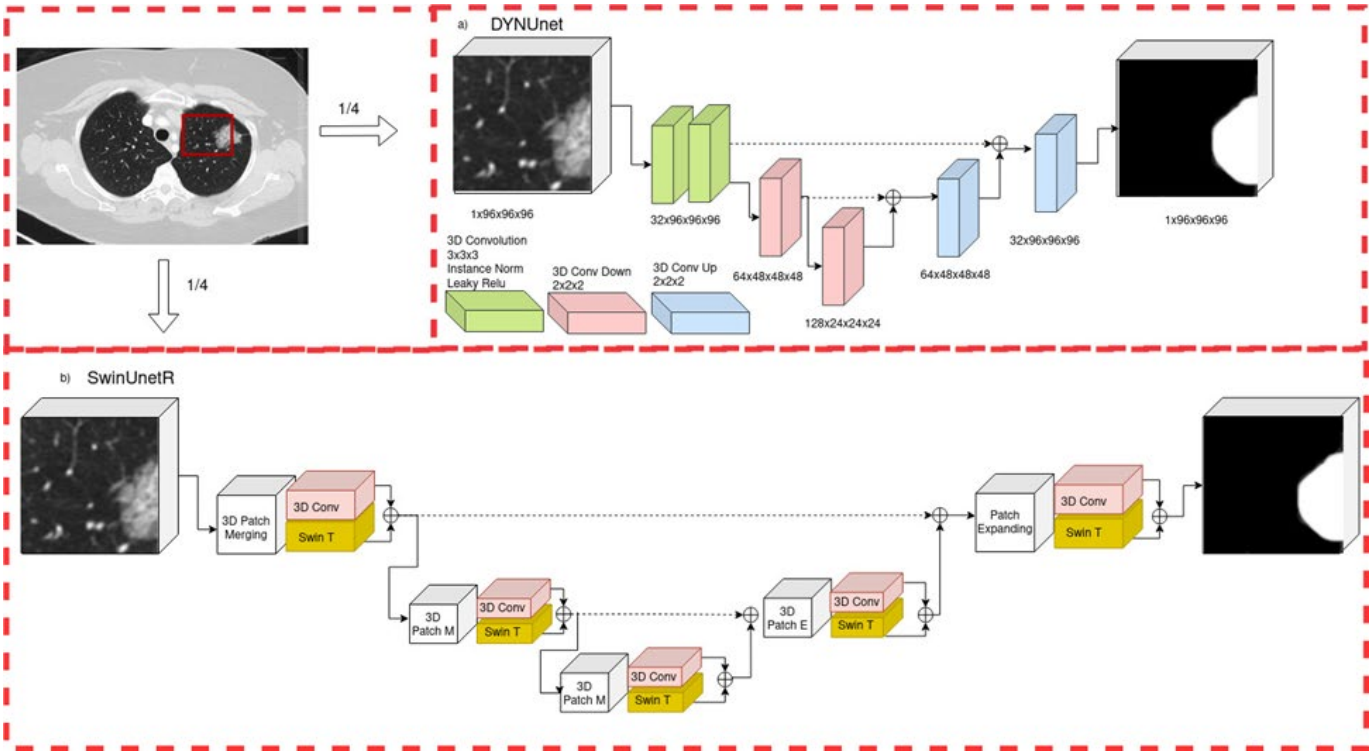


Illustration of both segmentation networks. (a) DYNUnet with downsampling dynamic strategy of two different kernel sizes. (b) SwinUnetR with Swin Transforms, as green blocks, attached to each layer of downsampling and upsampling.

Keywords

Deep Learning; Lung Cancer; SABR; Image Segmentation; 4DCT; Tumor Contouring

Advancing Hepatic Decompensation Status Prediction through Computed Tomography-Based Radiomics Signature and Body Composition Model Integration

Yashbir Singh, PhD, Assistant Professor, Mayo Clinic; John Eaton, MD; Sudhakar Venkatesh, MD; Bradley Erickson, MD, PhD, CIIP, FSIIM

Introduction

Hepatic decompensation is a critical manifestation of primary sclerosing cholangitis, a chronic cholestatic liver disease. Accurate prediction of hepatic decompensation status is crucial for timely intervention and management. In this study, we aimed to investigate the potential of Computed Tomography (CT)-based radiomics signature combined with the Body Composition Model to predict the occurrence of hepatic decompensation.

Hypothesis

We hypothesized that leveraging radiomics features extracted from the body composition compartments, as quantified by the Body Composition Model, could provide valuable insights into predicting hepatic decompensation status.

Methods

The Institutional Review Board approved this study and the informed consent procedure. The inclusion criteria were a) diagnosis of PSC and b) availability of an abdomen CT acquired during the portal venous phase. A total of 80 patients were included: 30 patients had hepatic decompensation, 30 had no hepatic decompensation, and 20 patients were part of an external validation set during 5-fold cross-validation. The study utilized the Body Composition Model, a deep learning framework developed in-house, which enables the quantification of four distinct compartments from CT scans: subcutaneous adipose tissue, skeletal muscle, visceral adipose tissue, and intermuscular adipose tissue. Radiomics features were extracted from each of these compartments, and a predictive model was constructed using a training cohort.

Results

The constructed predictive model demonstrated excellent performance in the validation cohorts for predicting hepatic decompensation status. It achieved an accuracy score of 0.97, precision score of 1.0, and an Area Under the Curve (AUC) score of 0.97, indicating its robustness and reliability in identifying patients at risk of hepatic decompensation.

Conclusion

The integration of CT-based radiomics signature with the Body Composition Model presents a promising approach for predicting hepatic decompensation status in patients with primary sclerosing cholangitis. The model's high accuracy and precision highlight its potential as a valuable tool in clinical practice for early identification of

individuals who may require intervention and proactive management.

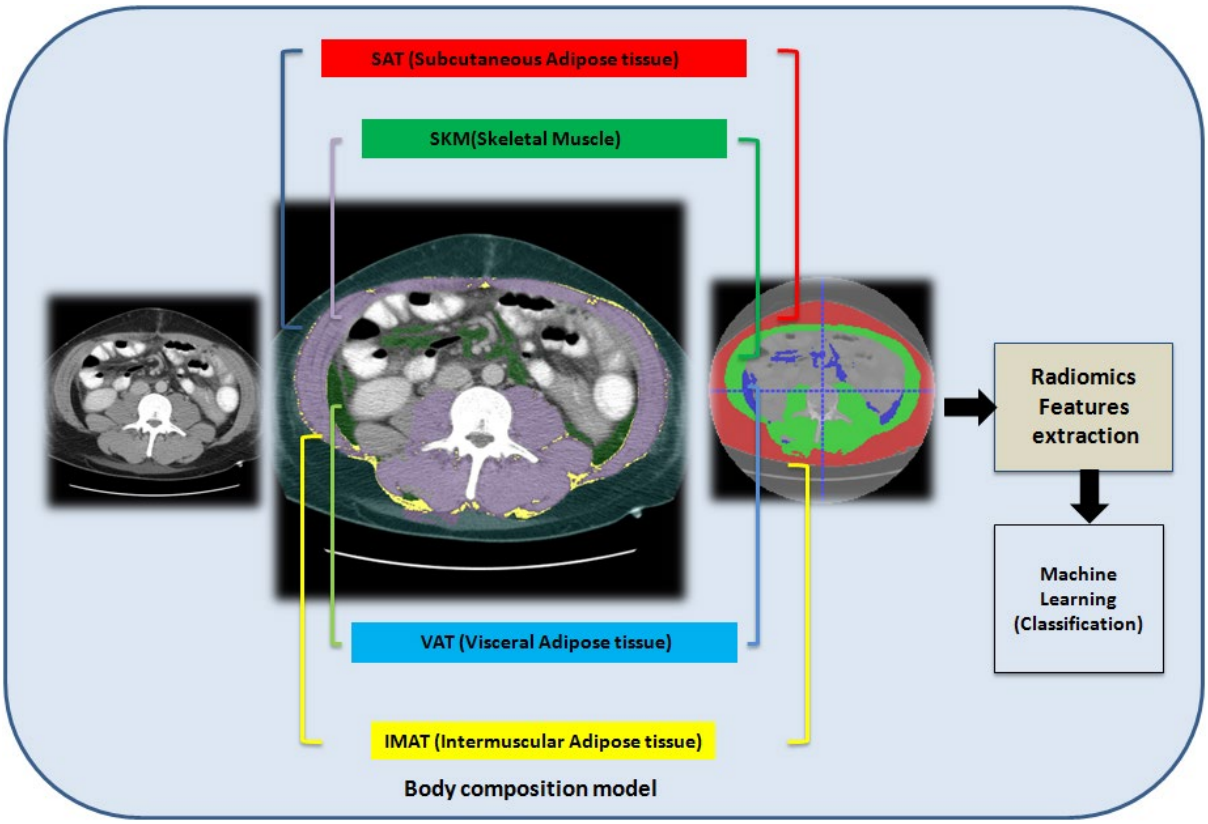
Statement of Impact

This study showcases the significant potential of incorporating advanced computational techniques, such as radiomics and deep learning, in predicting hepatic decompensation status. The accurate and early identification of patients at risk of hepatic decompensation can lead to timely interventions and improved outcomes, enhancing the overall management of primary sclerosing cholangitis patients. The findings of this study contribute to the growing body of knowledge in the field of precision medicine and highlight the value of integrating imaging-based biomarkers in clinical decision-making processes.

Table 1. Predicted values for metrics obtained during a 5-fold stratified cross-validation evaluation of the random forest classifier in the derivation cohort.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy score	0.94	0.94	0.87	1.0	0.98
Precision-score	0.89	0.94	0.83	1.0	1.0
Recall Score	1.0	0.93	0.93	1.0	0.94
AUC score	0.94	0.94	0.87	1.0	0.97

Workflow of the Prediction of Primary Sclerosis Cholangitis using Computed Tomography-Based Radiomics Signature and the Body Composition Model



Keywords

Radiomics; Body composition; Machine learning; Primary Sclerosing Cholangitis; Computer Tomography

An interactive Decision Support Tool for Evaluating Machine Learning Algorithm Performance in Medical Image Analysis developed by the Medical Imaging and Data Resource Center (MIDRC): MIDRC-MetricTree

Tingting Hu, PhD Candidate, Visiting Scientist, U.S. Food & Drug Administration; Natalie Baughan, PhD; Karen Drukker, PhD, MBA; Maryellen L. Giger, PhD; Grace Hyun Kim, PhD, MS; Mike McNitt-Gray, PhD; Berkman Sahiner, PhD; Emily Townley, Heather Whitney, PhD

Background/Problem to be solved

The Medical Imaging Data and Resource Center (MIDRC) was established to support medical imaging machine learning (ML) research for tasks including early detection, diagnosis, prognosis, and assessment of treatment response related to the COVID-19 pandemic and beyond. One specific project within MIDRC, the technology development project (TDP) 3c, focuses on creating a publicly available resource to assist researchers in selecting appropriate metrics for evaluating the performance of their medical imaging machine learning algorithms.

Intervention(s)

We have developed a task-oriented interactive decision tree, known as the MIDRC-MetricTree, which is structured according to the specific task for which the machine learning algorithm was trained. The decision tree allows users to select information such as the task type, nature of the reference standard nature (e.g. negligible unreliability and variability or non-negligible variability), and algorithm output type (e.g. binary or continuous data). Based on user inputs, the decision tree provides information resources on suitable performance evaluation approaches and metrics. These resources include literature references, links to publicly available software (when possible), and short tutorial videos.

Outcome

Five types of tasks were identified for the decision tree: (a) classification, (b) detection/localization, (c) segmentation, (d) time-to-event analysis, and (e) estimation. As an example, the classification branch of the decision tree covers binary and multi-class classification tasks and provides suggestions for methods, metrics, software links, and literature references for situations where the algorithm produces either binary or non-binary (e.g., continuous) output and for reference standards with negligible or non-negligible variability and unreliability.

Conclusion

The decision tree has been made publicly accessible (https://www.midrc.org/performance-metrics-decision-tree) and serves as a resource for researchers conducting task-specific performance evaluations in areas such as classification, detection/localization, segmentation, time-to-event analysis, and estimation. Our tree is expected to foster the development of AI/ML algorithms by helping users select appropriate metrics and by providing resources for applying selected metrics to their use cases for rigorous performance evaluation.

Statement of Impact

Selecting appropriate metrics for evaluating AI/ML algorithms can be challenging, particularly when faced with numerous options (e.g., for binary classification tasks) or a lack of standardized evaluation approaches (e.g., for complex multi-class classification tasks). Our work addresses this challenge by providing an interactive, user-friendly decision tree tool that offers suggestions to task-based performance evaluation approaches and metrics. While the decision tree is continuously evolving due to the dynamic nature of research in these areas, it serves as a valuable resource to help users narrow down the scope of metric selection in their search.

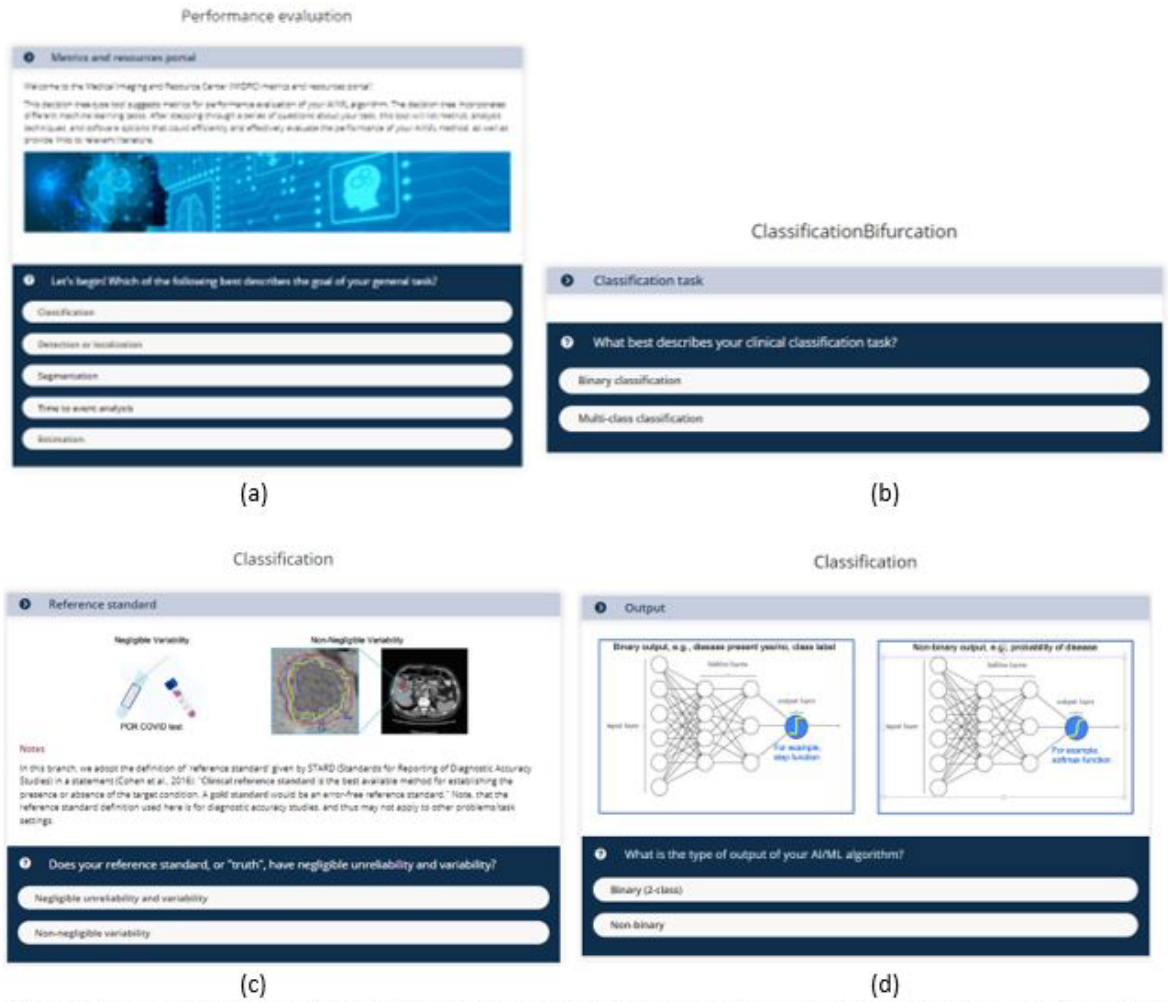


Figure 1: (a) The starting point of the interactive decision tree where the user selects the appropriate ML task; (b) For the classification task, the next question to be answered is whether the task is binary or multi-class classification; (c) In this example, the binary classification task was selected. Within this task, the next node asks the user to select between two options regarding the reference (or "truth") standard with regard to the variability of that standard; (d) the next node is where the user selects the type of output of the AI/ML algorithm (whether the algorithm is designed to output binary or non-binary (e.g. continuous) values).

Figure 1: (a) The starting point of the interactive decision tree where the user selects the appropriate ML task; (b) For the classification task, the next question to be answered is whether the task is binary or multi-class classification; (c) In this example, the binary classification task was selected. Within this task, the next node asks the user to select between two options regarding the reference (or "truth") standard with regard to the variability of that standard; (d) the next node is where the user selects the type of output of the AI/ML algorithm (whether the algorithm is designed to output binary or non-binary (e.g. continuous) values).

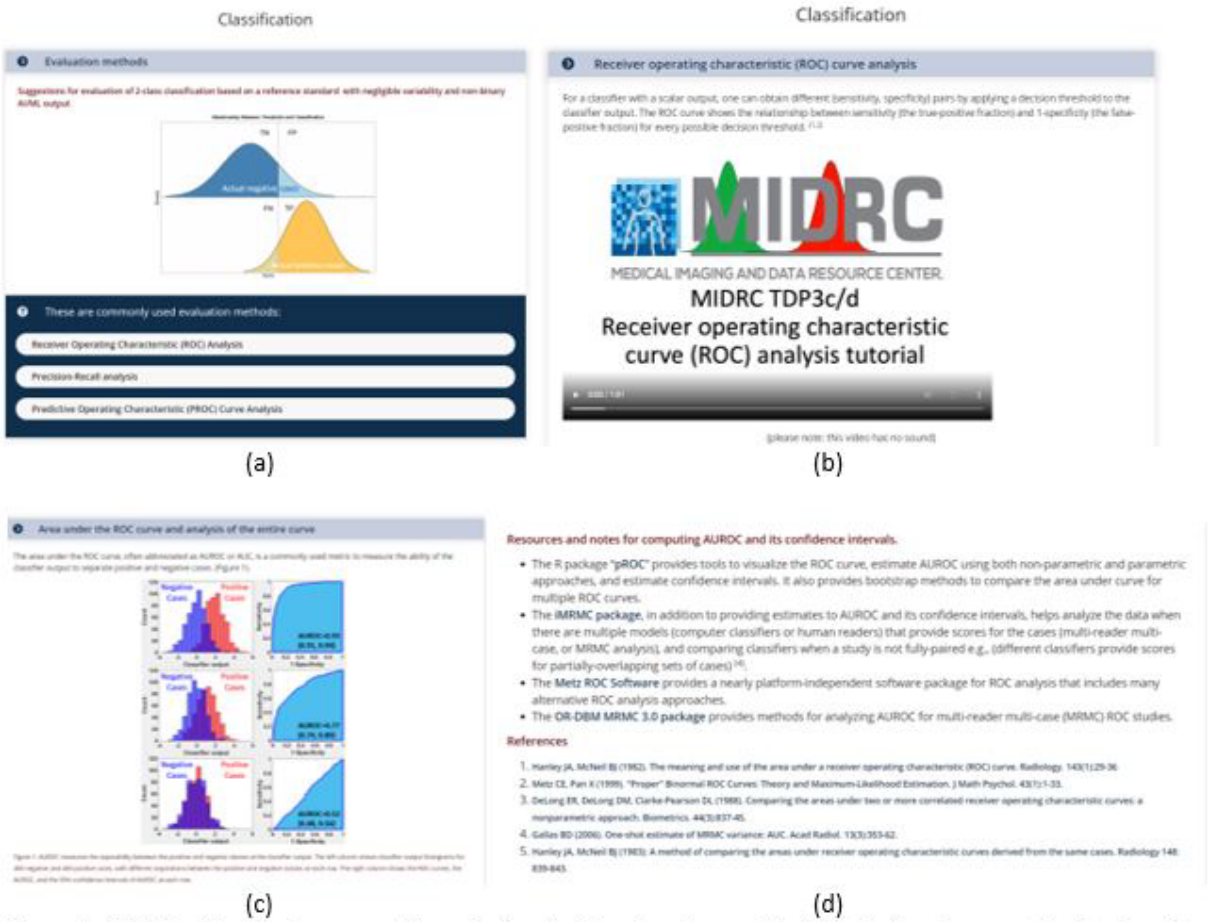


Figure 2: (a) After the user has gone through the decision tree to specify their task, reference standard, and type of machine learning output, the decision tree suggests one or more evaluation approaches. Based on the user's responses, the decision tree may provide (b) a tutorial (here showing one on ROC analysis); (c) details on specific metrics to be used in the performance evaluation (here showing area under the ROC curve; and (d) a list of vetted resources such as software packages that compute the desired metric (preferably including error estimates) as well as suggested literature references.

Figure 2: (a) After the user has gone through the decision tree to specify their task, reference standard, and type of machine learning output, the decision tree suggests one or more evaluation approaches. Based on the user's responses, the decision tree may provide (b) a tutorial (here showing one on ROC analysis); (c) details on specific metrics to be used in the performance evaluation (here showing area under the ROC curve; and (d) a list of vetted resources such as software packages that compute the desired metric (preferably including error estimates) as well as suggested literature references.

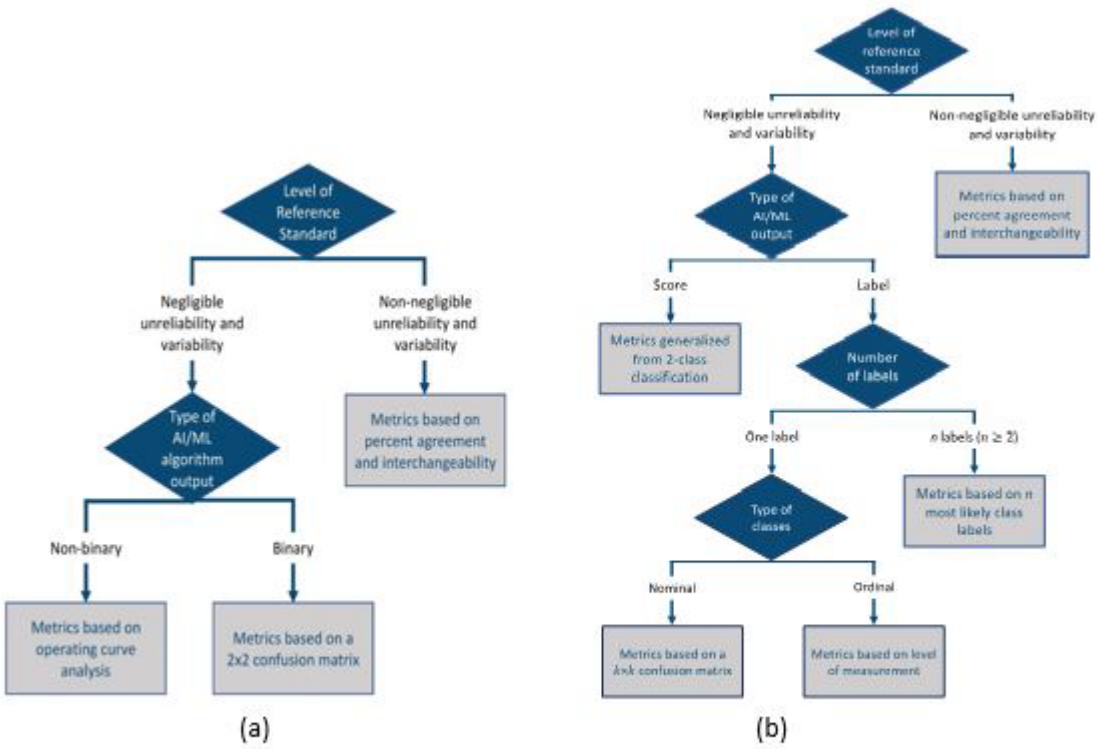


Figure 3: (a) A simplified flowchart for the Binary Classification Branch of the MIDRC-MetricTree (b) A simplified flowchart for the Multi-class Classification Branch of the MIDRC-MetricTree

Figure 3: (a) A simplified flowchart for the Binary Classification Branch of the MIDRC-MetricTree (b) A simplified flowchart for the Multi-class Classification Branch of the MIDRC-MetricTree

Keywords

Artificial Intelligence; Machine Learning; Computer-Aided Diagnosis; Performance Evaluation

Analysis of Intersectional Bias in a Novel Melanoma Image Classification Algorithm

Christopher Caligiuri, Student, Princeton University

Background/Problem to be solved

With the rising incidence of melanoma worldwide, it is important to ensure that machine learning (ML) algorithms developed to classify suspicious skin lesions are fair and distinguish melanoma lesions in all skin types. Existing research has explored bias in a variety of machine learning applications in healthcare (Larrazabal, A. et. al, 2020). No attempts have been made to determine the consistency in performance of melanoma classification (ML) algorithms across gender and varying skin tones.

Intervention(s)

To evaluate the efficacy of the deep convolutional neural network across genders and skin tones, an algorithm released by Soenksen et al. on GitHub was interfaced. The code was executed on Jupyter notebook and modified to accommodate individual images. Each image was classified as male/female, a Fitzpatrick skin type, and as malignant/benign based on the algorithm's prediction. Of the 33,126 total images, 700 dermoscopic images were randomly selected from the database for testing with the algorithm. To assign a Fitzpatrick skin type, a novel Python application for determining skin type was developed. The application was implemented to ensure that the pigmented lesions were not used to determine the skin type. The average RGB values were then extracted from the selected region, and the corresponding Fitzpatrick skin type was outputted. The Fitzpatrick skin type was based on the RGB averages for each skin type as reported by a team of dermatologists (Jo and Kim, 2019).The randomly selected skin lesions were inputted into the DCNN. The ground truth diagnoses provided by the SIIM-ISIC database were compared against predicted values.

Outcome

The analysis revealed statistically significant skin tone bias in the DCNN algorithm, with variations in accuracy between different Fitzpatrick skin types. The maximum difference in accuracy was found to be 0.33 between skin type II and V, indicating a potential for disparities in melanoma classification based on skin tone. However, minimal gender bias was observed in the algorithm's performance.

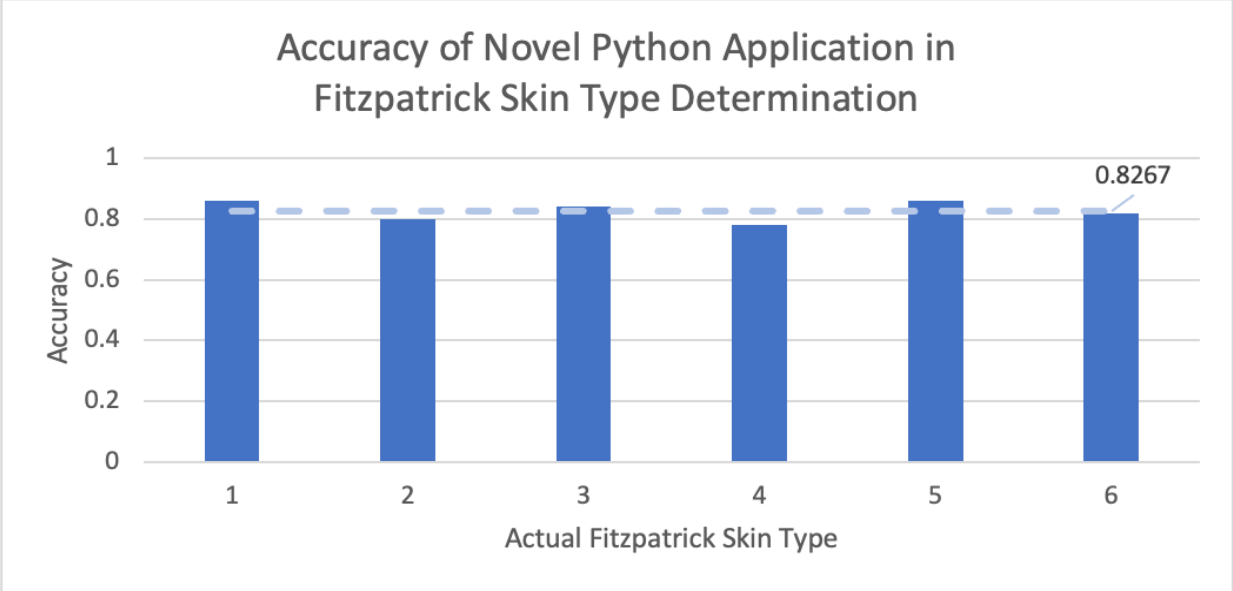
Conclusion

The study demonstrates the existence of intersectional bias in a novel melanoma image classification algorithm. The findings highlight the importance of considering and addressing bias in ML algorithms used for healthcare applications.

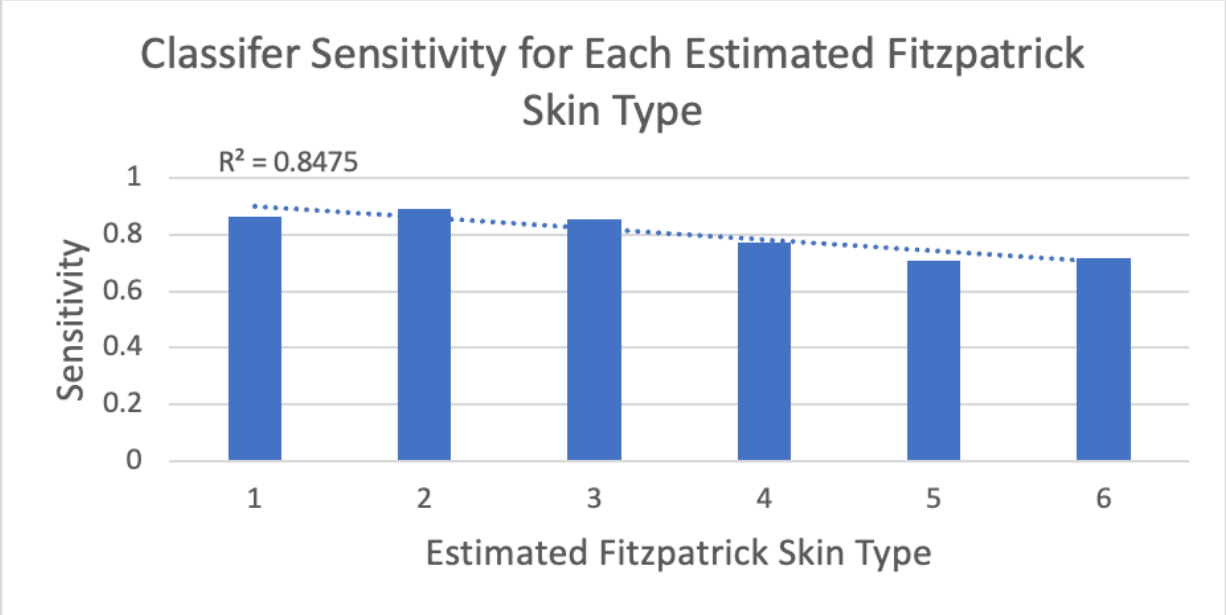
Statement of Impact

This research improves AI-based melanoma diagnosis in diverse populations by addressing bias in ML algorithms,

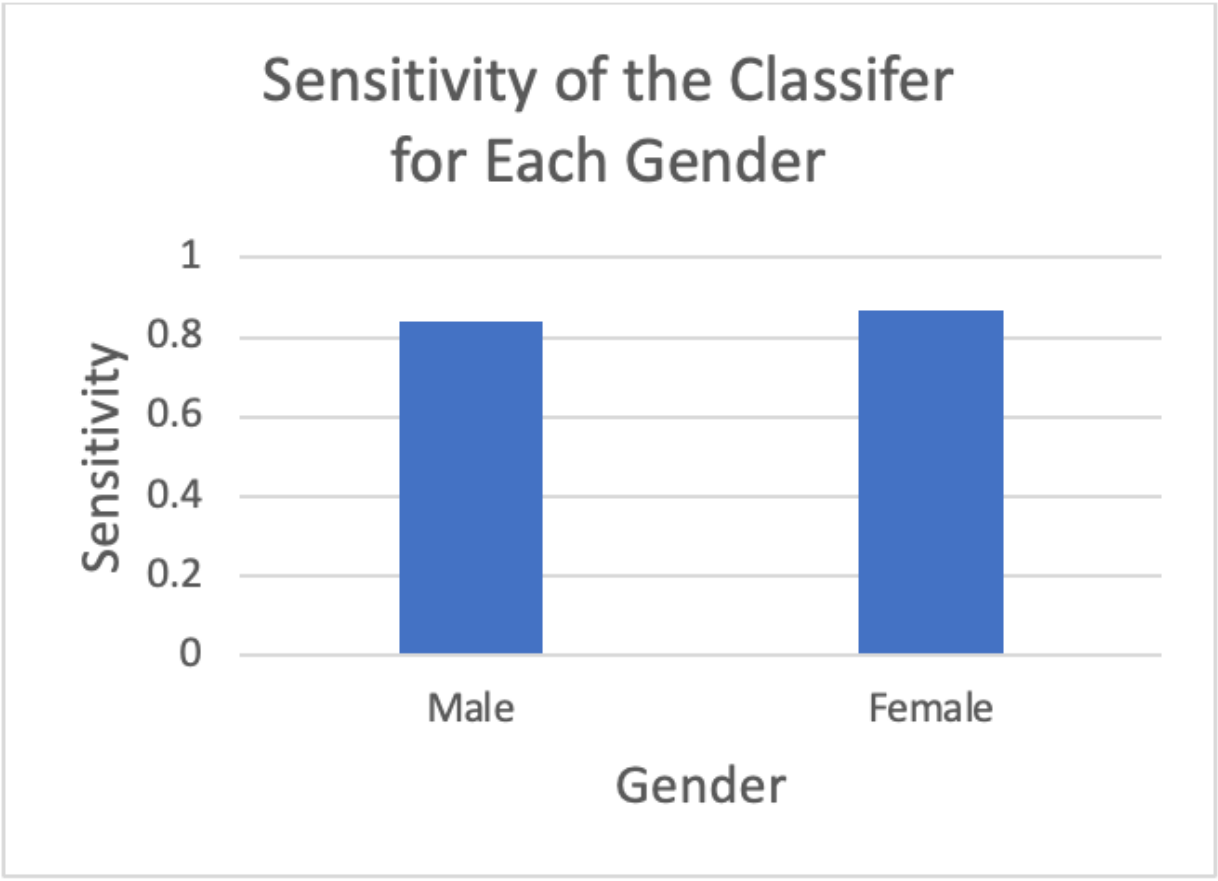
ensuring equitable screening and better outcomes. Further, the approach could be applied in other settings to simplify the determination of Fitzpatrick skin types and allow researchers to incorporate racial analysis more accurately.



The accuracy of the Python application in estimating the Fitzpatrick skin type. There is a relatively high average of 82.67%, qualifying its use in the study.



The sensitivity of the classifier for each skin type along with a linear trendline overlayed over the histogram. As indicated by the trendline, the model seems to decrease in performance with darker skin types. There is also a significant drop between skin type III and IV, with V having the worst accuracy.



The distribution of the sensitivity of each gender in the selected images from the SIIM-ISIC database. There seems to be little difference in the male and female distributions.

Keywords
Melanoma; diagnosis; Fitzpatrick skin types; image classification; bias

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

ChatGPT Enhanced Radiology Reporting using PRECISE Framework For Patient-Centered Care

Satvik Tripathi, Research Student, Massachusetts General Hospital; Emiliano Garza, MD; Liam Mutter; Michael Dezube, MEng; Christopher P. Bridge, PhD; Dania Daye, MD, PhD

Introduction
Radiology reports play a crucial role in medical diagnostics, providing valuable insights and interpretations of medical imaging studies to guide patient care. However, these reports can be challenging for patients to understand due to their technical language and lack of patient-centered information. In this research project, we explore the application of ChatGPT, a state-of-the-art language model, to enhance radiology reporting and propose patient-centered text using the PRECISE (Patient-Focused Radiology Reports with Enhanced Clarity and Informative Summaries for Effective Communication) framework. By incorporating patient-centric principles, such as relevance, explicitness, conciseness, and effectiveness, we aim to revolutionize patient-centered care in radiology.

Hypothesis
We propose the PRECISE framework, derived from radiology reports utilizing ChatGPT. These PRECISE texts are patient-friendly and can be understood without the need for any medical domain knowledge.

Methods
Utilizing a publicly available chest X-ray dataset from Indiana University, we generated patient-friendly and readable text summarizing radiology reports by utilizing the findings section of 200 reports and ChatGPT (GPT-3.5) with the prompt "summarize the given radiology report." We conducted three assessments on the PRECISE text: Readability, Reliability, and Explainability. Readability was evaluated with standard scores (Flesch Reading Ease, Gunning Fog Index, Automated Readability Index). For Reliability, an expert clinician assessed the medical correctness, categorizing it as "Appropriate," "Inappropriate," or "Unreliable." For Explainability, a non-medical volunteer rated the text on a scale of 0-2 (0 = not explainable, 1 = well explained with some terms left unexplained, 2 = everything explained). All assessments were conducted independently, with graders only seeing the report and corresponding PRECISE text.

Results
The PRECISE text performed well in all assessments. In the readability tests, the average scores were 55±13.2 (Flesch Reading Ease), 11±2.7 (Gunning Fog Index), and 11±2.6 (ARI), indicating readability at an 11th-grade level or above. In the reliability test, 92% were classified as "appropriate," 8% as "inappropriate," and none as "unreliable" or containing false medical information. In the explainability test, 7% were class 0, 23% class 1, and 70% class 2.

Conclusion
Our study demonstrates the significant potential of the PRECISE framework in transforming radiology reporting into a patient-centered approach. The PRECISE text's high readability and reliability scores suggest it balances clear communication with medical information match expert clinician assessments. The explainability test reinforces the effectiveness in non-medical readers.

Statement of Impact
This approach empowers patients with a better understanding of their medical condition, facilitating informed discussions with healthcare providers and fostering active participation in their own care.

Figure. 1. An illustration of the PRECISE framework in clinical workflow

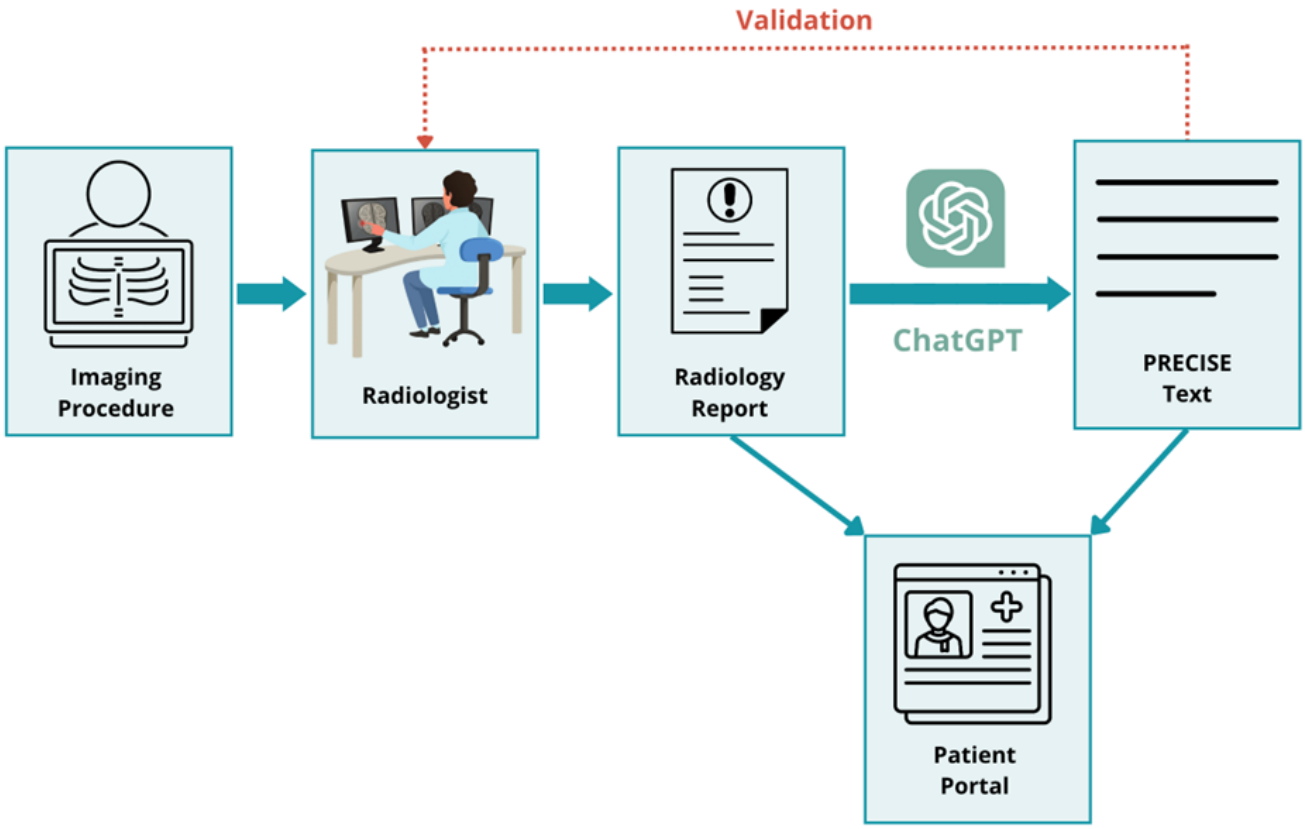
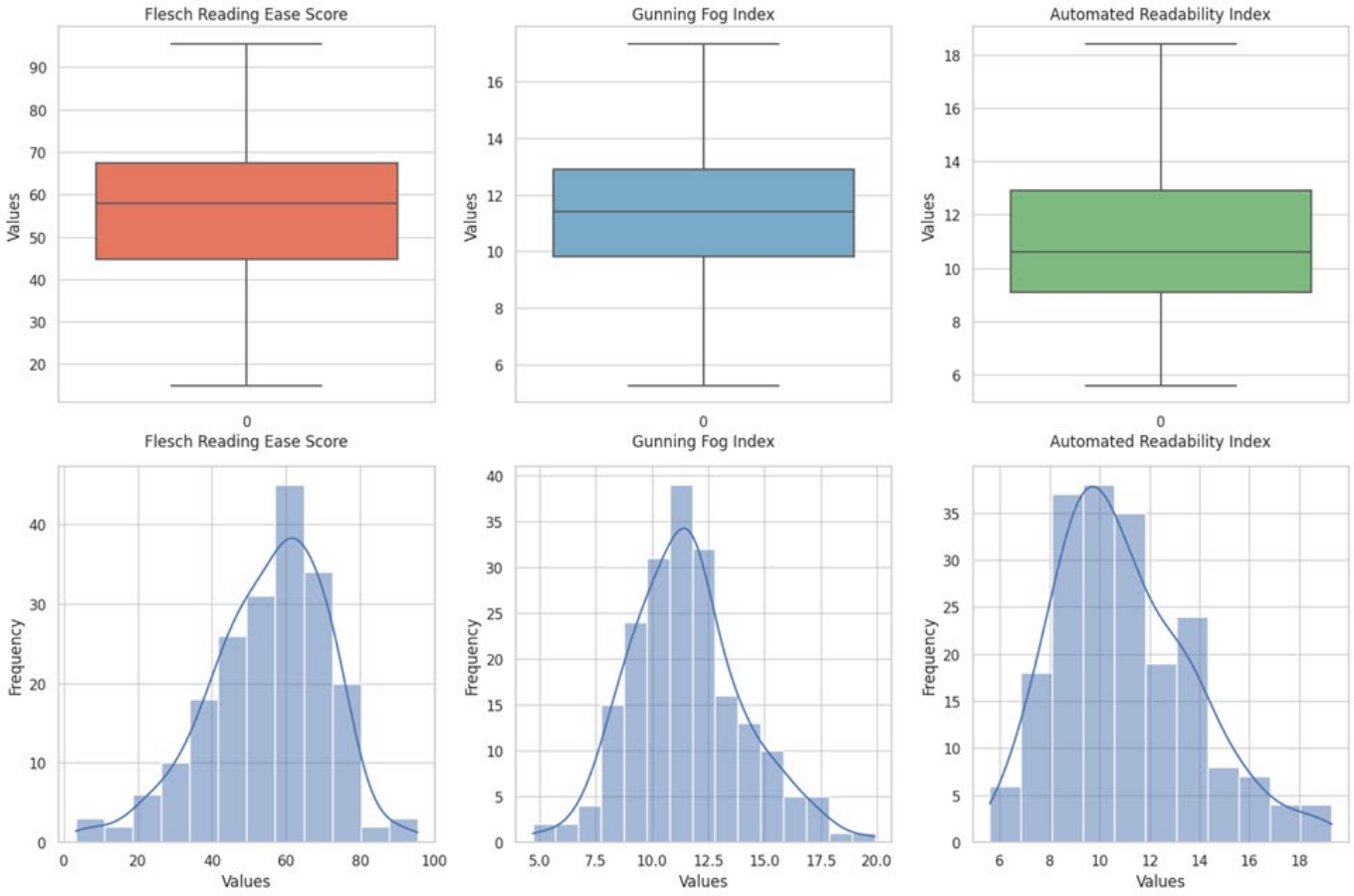


Figure. 2. Distribution of Flesch Reading Ease Score, Gunning Fog Index, and ARI score



Keywords
Generative AI; Artificial Intelligence; ChatGPT; Patient-centered Care; Radiology Reports

Deep Learning-Based Natural Language Processing for Classification of Renal Surgical Pathology Outcomes in a Multi-Site Dataset

Satvik Tripathi, Undergraduate Researcher, Perelman School of Medicine at the University of Pennsylvania; Rithvik Sukumaran; Jackson Steinkamp; Charles M. Chambers; Darco Lalevic; Hanna M. Zafar, MD, MHS; Tessa S. Cook, MD, PhD, CIIP, FSIIM, FCPP

Background/Problem to be solved

Accurate classification of pathology is essential for guiding clinical care and measuring patient outcomes. Natural language processing (NLP) can automate the analysis of medical data, including pathology reports. In this study, we developed and evaluated a deep learning-based NLP framework on a multi-site database to accurately categorizing classify renal surgical pathology outcomes as benign, indeterminate, or malignant.

Intervention(s)

We utilized a multi-channel convolutional neural network (CNN) model for the classification task. The model consists of a single convolutional layer followed by a dropout layer, max pooling, and a fully connected layer for classification. The model was trained on 688 renal surgical pathology reports from three different health systems in the state. The reports were labeled by two radiologists who interpret abdominal imaging, with 10 and 14 years of experience, respectively. The four labels used were: “malignant”, “indeterminate”, “benign”, and “ignore” (n = 140, 51, 192, and 305 respectively). “Ignore” was used to indicate cytopathology, urinalysis or urine cultures, or any other pathology not specifically from a mass lesion in the kidney. Only the summary text from each pathology report was used and tokenized. The model was validated using a 20% holdout validation sample using 10-fold cross-validation and evaluated using cross-entropy loss.

Outcome

We achieved an average validation accuracy of 83.67% across all the folds. The model was also evaluated on precision, recall, and F1 score. The model is generalized and robust over all three sites and does not exhibit any site bias. The model performance was best in the "ignore" class while suffering in the "indeterminate" class due to class imbalances in the dataset. Performance in the classification of “malignant” and “benign” classes remained consistent through each fold.

Conclusion

The proposed framework will not only improve the efficiency and consistency of renal pathology assessment but also have the potential to uncover novel insights and correlations within the dataset. We have developed a highly accurate NLP framework that automates the classification of renal surgical pathology outcomes. These automated pathology data can be correlated with radiology reports in order to map imaging features with patient outcomes and provide educational feedback to radiologists.

Statement of Impact

Automated labeling of surgical pathology reports can enhance clinical decision-making, automation, patient management, and the overall understanding of renal diseases.

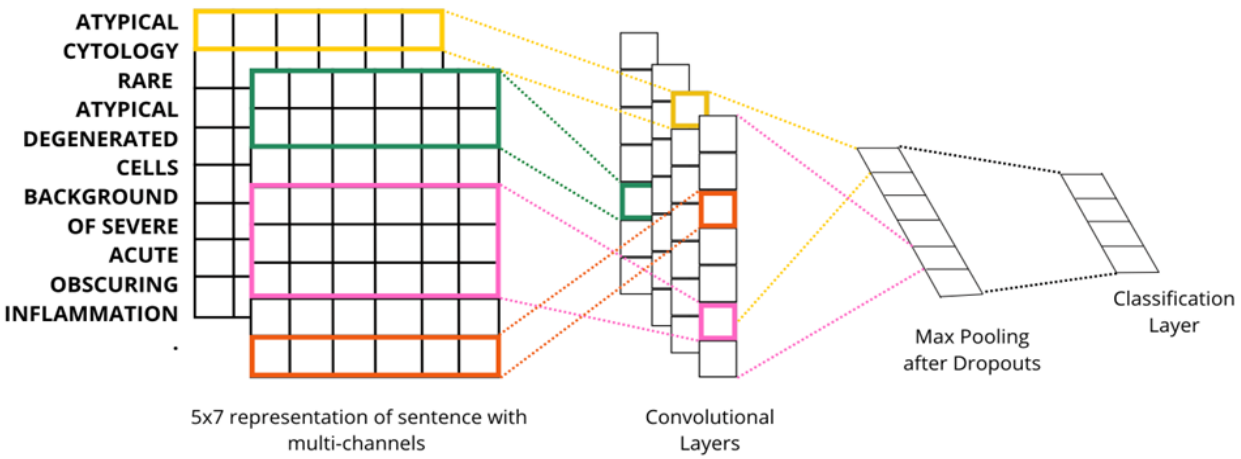


Figure.1. Model architecture of the multi-channel CNN used in this study.

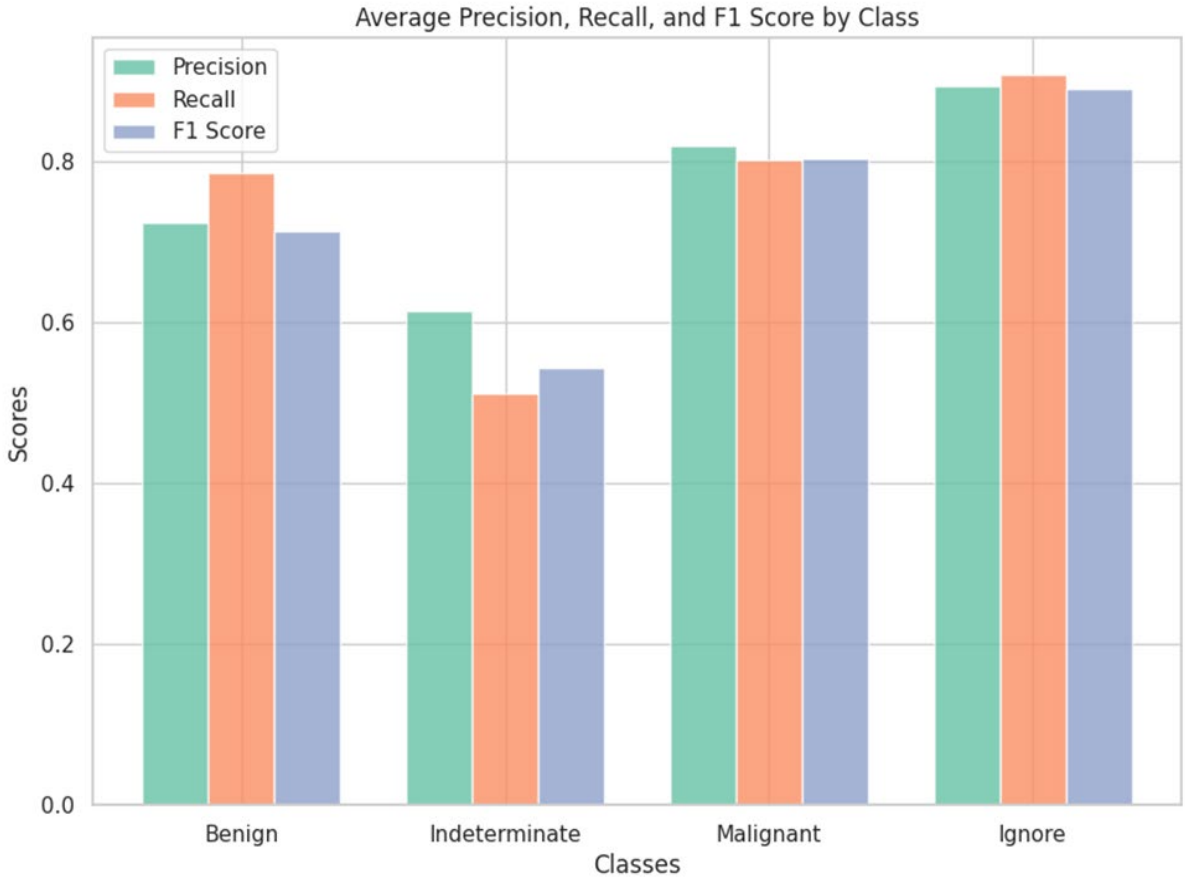


Figure.2. Average precision, recall, and F1 score for each class across all folds

Keywords
Deep learning; Natural language processing; Cancer; Pathology

Deep Learning Assisted Curation of the CANDID-III Dataset with Free-text Reports

Anna Hu, Medical Student, George Washington University School of Medicine and Health Sciences;
Sijing Feng, MBChB; Qixiu Liu, MBChB; Darren Ritchie, MBChB; Benjamin Wilson, MBChB, FRANZC

Background/Problem to be solved

This project aims to curate the CANDID-III dataset, which consists of adult chest radiographs with comprehensive labels derived from both manual and AI-assisted annotation.

Intervention(s)

The CANDID-II dataset is an in-development chest radiograph dataset containing 33,486 anonymized free-text radiological reports. CANDID-III inherited the same 45 radiological labels from the CANDID-II dataset, which were mapped to UMLS ontology for standardization, forming the manually labelled portion of the CANDID-III dataset. An ensemble transformer-based label extraction model, combining three individual natural-language processing (NLP) algorithms, was trained and validated on the CANDID-II dataset in an 80:20 proportion. Each algorithm was individually trained on every radiological label, and the highest accuracy algorithm was chosen on a per-label basis for inclusion into the final ensemble model. The model was then used to automatically label the remaining CANDID-III dataset. An evaluation set of 552 reports, with balanced sampling across radiological findings from the AI-labeled portion of the CANDID-III dataset, was assessed by selected annotation team members, including a final-year radiology trainee and a fourth-year postgraduate medical doctor. Label-specific ‘mention’ F1 scores were calculated for the final ensemble model, with ‘not mentioned’ as negative and ‘indeterminate, absent, present’ as combined positive classifications.

Outcome

The completed CANDID-III dataset contains 322,473 images and 220,977 anonymized free-text radiological reports from 94,210 unique patients (1:1.04 M:F ratio). AI-assisted annotation was performed on 88% of the CANDID-III dataset. For the AI-assisted annotation portion of the CANDID-III dataset, the labelling model has a macro-F1 score of 0.88 and micro-F1 score of 0.94 across all findings. Seven labels are shared with CheXpert, with F1 scores ranging from 0.93 to 1.0. F1 scores for 30 CANDID-III labels are above 0.90, while 8 labels range between 0.80 and 0.90.

Conclusion

The CANDID-III dataset provides a large, comprehensively labeled, and high-quality adult chest radiograph dataset with anonymized free text reports. The dataset adds numerous new clinically significant radiological annotations that are labelled to a high accuracy and contributes to the repertoire of publicly available chest radiograph datasets for AI development.

Statement of Impact

The CANDID-III dataset can be used to train and test AI algorithms for a variety of applications including triaging, lung cancer screening, image generation, and automated preliminary detection of radiographic abnormalities.

Table 1. Accuracy of AI-assisted annotation of the CANDID-III dataset.

	Mention F1 Score		
Radiological feature label	**CANDID**	**CheXpert**	**NIH**
Cardiomegaly	0.998	0.973	0.647
Pleural effusion	1.000	0.996	0.985
Consolidation	0.978	0.999	0.996
Pneumothorax or hydropneumothorax	0.993	1.000	0.993
Atelectasis	0.956	0.998	0.976
Fracture	0.930	0.975	N/A
Lung opacity	0.961	0.966	N/A
Pleural calcification	1.000	N/A	N/A
Surgical emphysema	1.000	N/A	N/A
Mastectomy	1.000	N/A	N/A
Pectus excavatum	1.000	N/A	N/A
Lung hyperinflation	0.990	N/A	N/A
Abnormal aortic morphology	0.977	N/A	N/A
Subdiaphragmatic free air	0.973	N/A	N/A
Bronchiectasis	0.968	N/A	N/A
Pulmonary cavitary lesion	0.960	N/A	N/A
Perihilar opacity	0.959	N/A	N/A
Pleural thickening	0.958	N/A	N/A
Lung nodule/mass	0.942	N/A	N/A
Oesophageal dilatation	0.941	N/A	N/A
Pulmonary fibrosis	0.933	N/A	N/A
Increased interstitial markings	0.927	N/A	N/A
Bone degenerative changes	0.926	N/A	N/A
Elevated hemidiaphragm	0.923	N/A	N/A
Hiatus hernia	0.919	N/A	N/A
Pneumomediastinum	0.917	N/A	N/A
Decreased interstitial markings	0.909	N/A	N/A
Bronchial wall thickening	0.909	N/A	N/A
Retrocardiac opacity	0.900	N/A	N/A
Tracheal shift	0.900	N/A	N/A

Mediastinal mass	0.894	N/A	N/A
Costophrenic blunting	0.889	N/A	N/A
Hilar enlargement	0.857	N/A	N/A
Right paratracheal soft tissue prominence	0.857	N/A	N/A
Non-surgical foreign body	0.857	N/A	N/A
Diaphragmatic hernia	0.833	N/A	N/A
Small lung volume	0.800	N/A	N/A
Gallstone	0.800	N/A	N/A
Device malposition	0.777	N/A	N/A
Tracheal stenosis	0.727	N/A	N/A
Bone lesion	0.714	N/A	N/A
Bony Metastases	0.667	N/A	N/A
Eventration of diaphragm	0.667	N/A	N/A
Calcified mediastinal lymph nodes	0.667	N/A	N/A
Axillary calcified lymph nodes	0.667	N/A	N/A
Macro-average	0.876	N/A	N/A
Micro-average	0.944	N/A	N/A

Keywords

Conventional Radiography; Machine Learning; Artificial Intelligence; Dataset

Evaluation of ChatGPT Performance on Radiology Board Exam-Style Questions

Anna Hu, Medical Student, George Washington University School of Medicine and Health Sciences; Sijing Feng, MBChB; John Egbuji, MBChB; Sophia Dean, MBChB; Ben K. Wilson, MBChB

Background/Problem to be solved

The American College of Radiology (ACR)'s clinical radiology examinations represent a major barrier to training progression for residents in training. ChatGPT is an emerging technology which has been shown to have the potential to help with medical education. Currently, there are 2 models, GPT-3.5 and GPT-4, the latter of which has been previously shown to perform better in reasoning tasks on professional and academic benchmarks compared to its predecessor. However, the usefulness of ChatGPT as a tool for preparing for radiology exams and whether ChatGPT can perform better given an exam-targeted knowledge base remains to be determined. This study aims to determine the performance of ChatGPT on ACR-style text-based multiple choice questions (MCQ) and to explore any improvement in the performance following the provision of exam-targeted knowledge base.

Intervention(s)

ACR text-based questions (40 pathology and 40 radiodiagnosis MCQs) were constructed by 4 board-certified radiology residents and attendings. The questions were reviewed and further subdivided into higher order and lower order questions according to Bloom's taxonomy. These questions were then tested on GPT-3.5 and GPT-4 models with and without the provision of an exam-targeted knowledge base. Statistical significance was determined by Chi-square analysis.

Outcome

For pathology MCQ, GPT-4 answered 92.5% of question correctly and outperformed GPT-3, which answered 57.5% of questions correctly ($P< 0.001$). GPT-4 also outperformed GPT-3 for radiodiagnosis MCQ (67.5% vs 42.5%; $P=0.025$). GPT-4 scored statistically significantly higher than GPT-3.5 on lower and higher-order questions for pathology (50% vs 80%, $P=0.047$ (low), 50% vs 85%, $P=0.018$ (high)), but such an improvement did not reach a statistical significance for radiodiagnosis questions (30.4% vs 52.2%, $P=0.13$ (low), 58.8% vs 70.6%, $P=0.47$ (high)). No significant difference was observed for both model with the addition of the knowledge base on either sub exam or between higher and lower-order thinking questions.

Conclusion

GPT-4 performed statistically significantly better than GPT-3.5 on both ACR radiology board exam-style pathology and radiodiagnosis MCQ. In addition, GPT-4 performed statistically significantly better on the higher and lower order pathology MCQ compared to GPT-3.5. Performance, however, did not improve given an exam-targeted knowledge base.

Statement of Impact

Without being specifically trained on clinical radiology corpus, GPT-4 performed significantly better than GPT-3.5 for both ACR board style pathology and radiodiagnosis MCQs. GPT-4 passed ACR radiology board exam-style pathology and radiodiagnosis MCQ. Such a high-quality performance from the GPT-4 model paves the way for potential future incorporation of large language models like ChatGPT into radiology training programs across all levels.

Table 1. Performance of GPT-4 and GPT-3.5 on ACR Radiology Board–style MCQ. Numbers in parentheses represent percentages.

		No. of Correct Responses		
Question Type	No. of Questions	GPT-3.5	GPT-4	P value
Pathology				
Baseline	40	23 (57.5)	37 (92.5)	<0.001
With knowledge base	40	28 (70)	38 (95)	0.003
Question type				
Lower-order				
Baseline	20	10 (50)	16 (80)	0.047
With knowledge base	20	11 (55)	16 (80)	0.05
Higher-order				
Baseline	20	10 (50)	17 (85)	0.018
With knowledge base	20	14 (70)	18 (90)	0.11
Radiodiagnosis				
Baseline	40	17 (42.5)	27 (67.5)	0.025
With knowledge base	40	18 (45)	30 (75)	0.006
Question type				
Lower-order				
Baseline	23	7 (30.4)	12 (52.2)	0.13
With knowledge base	23	5 (21.7)	12 (52.2)	0.031
Higher-order				
Baseline	17	10 (58.8)	12 (70.6)	0.47
With knowledge base	17	12 (70.6)	14 (82.4)	0.42

Keywords

natural language processing; Generative AI; radiology education; ChatGPT

Decoder-Only Computed Tomography Radiology Reports (DOCTRR)

Tegan Keigher, MS, Graduate Student, Data Science Institute, University of Chicago; Benan Akca, PhD; Andrew Alvarez, MS; Mary Erikson, MS; Utku Pamuksuz, PhD

Background/Problem to be solved

Significant advancements have been made in the realm of large language models (LLMs) within the healthcare industry. Previous research proposed impression generation by using the findings section and patient background as input (i.e. Gundogdu et al, 2021). However, generating accurate findings sections for CT radiology reports remains a challenging task that demands substantial time and effort from radiologists. This research addresses this challenge by harnessing the potential of LLMs to automatically generate findings sections.

Intervention(s)

We employ hybrid fine/instruct-tuning techniques to generate findings based on previously generated structured image labels. We designed a vector-to-sequence model capable of producing radiology reports with a level of accuracy and quality comparable to that of human physicians. Additionally, the research contributes to identifying the necessary image labels/ontology for comprehensive and precise findings generation. By incorporating structured image labels with a proper ontology that encompasses observations, locations, sizes, severities, clinical findings, and more, we input this information into a transformer model—specifically a pre-trained large language model (open-source, Alpaca and Falcon-40B). The model is fine-tuned using two different feature engineering methods on the training set.

Outcome

Encouraging results are observed in zero-, one-, and few-shot scenarios, and we anticipate further improvements through iterative tuning. We plan to conduct ablation studies with other open-source LLM's including MedPalm2 (Singhal et al., 2023) and Vicuna (Peng et al., 2023). We are testing our model in a real-time experimental setup with radiologists in a top-tier academic institution and statistically validating the performance using ROUGE metrics. We also incorporated Radlex and Snomed ontologies, applying text processing techniques to accurately structure the input. Preliminary observations have shown that our predictions align with the original free-form text. Figure 1 demonstrates sample results from three different techniques.

Conclusion

The proposed methodology presents promising initial results (59% Rouge Score Accuracy with 78% clinical validation) in revolutionizing the radiology workflow by automating the generation of findings sections. We draw on NLP and ontology methods to better assess the potential of our tool in an ideal scenario where both input labels and text generation are driven by AI-powered image processing techniques.

Statement of Impact

Our vision is to solve the text generation task as part of a fully automated radiology process that integrates image processing and text generation in a unified pipeline in the future. Based on our initial results and methodology, we are confident that this work will make a significant contribution towards achieving full automation in the future.

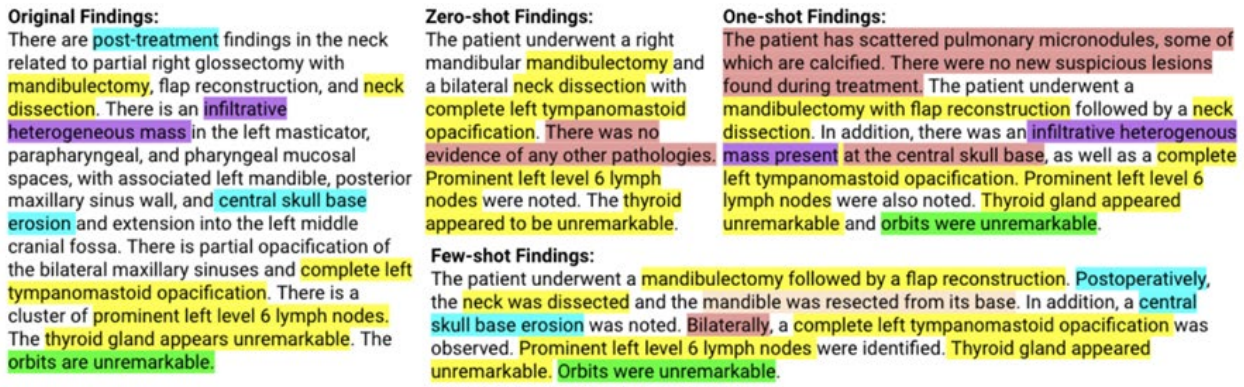
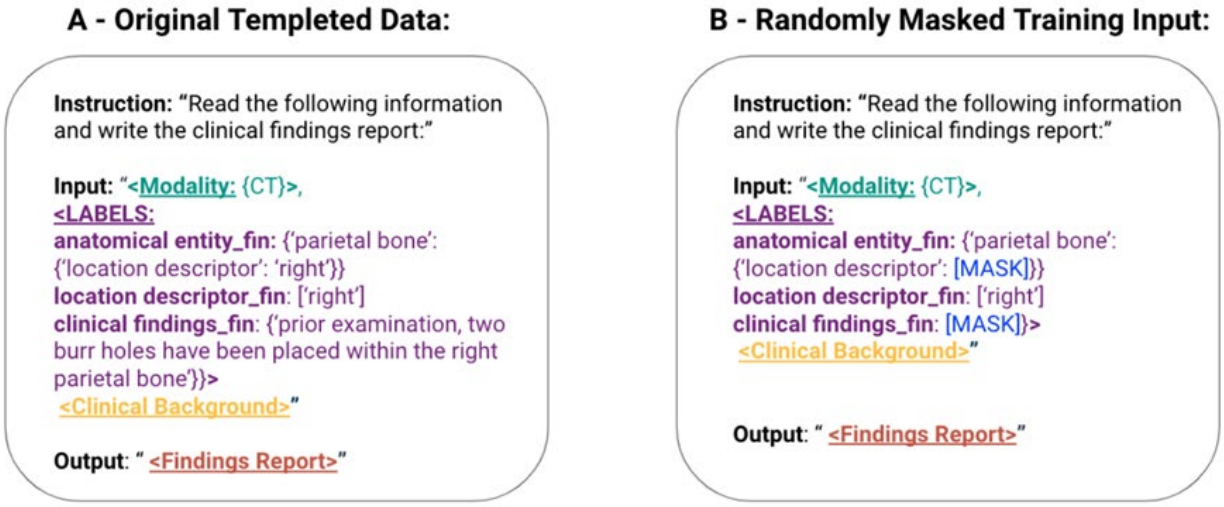


Figure 1. Results from zero-, one-, and few-shot testing. Text highlighted in yellow indicates that all three tests included this information. Text highlighted in purple indicates that only one-shot testing included this information. Text highlighted in green indicates that the information was included in both one- and few-shot attempts. Text highlighted in blue indicates that the information was included only in the few-shot attempt. Text highlighted in red indicates model hallucination, while text highlighted in orange indicates possible hallucination



Annotated Image Labels and Masking Process to detect the necessary image labels for comprehensive and precise findings generation.

Example of the entire minimum label requirement testing process. Each box represents a dataset with the corresponding label dropped from the input. Green indicates the average cosine similarity score was above the threshold and therefore should continue, red indicates it does not pass the threshold and to deem it a critical label combination

Keywords
LLM; Radiology; Findings; Generative AI; Ontology; Auto-generated Radiology Reports

Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension

Jen Park, MS, PhD Student, Johns Hopkins University; Kyulee Jeon; Haridimos Kondylakis, PhD; Teri Sippel Schmidt, MS, FSIIM; Paul Nagy, PhD, CIIP, FSIIM; Seng Chan You, MD, PhD

Background/Problem to be solved

Our goal is to combine clinical outcomes in the EHR with imaging studies, to define cohorts consistently, and to enable federated learning. The Observational Health Data Sciences and Informatics (OHDSI) open science collaborative is an international community that collaboratively generates evidence to promote better health decisions and better care. OHDSI uses an open community standards-based common data model for conducting reproducible real-world analyses on observational health data using open-source analytics software. Observational health data includes electronic health records (EHR) and administrative claims. OHDSI has developed processes to enable securely sharing cohorts and results across a distributed network for robust real-world evidence without sharing any patient-level data. As of 2022, over 450 health systems and claims databases have adopted the OMOP data standard, with medical records from over 960 million unique patients represented. This is approximately 1/8 of the world's population. Participation in this extensive data network presents an excellent opportunity for the imaging community. We have linked algorithmically generated measurements into the OMOP data model to harness these deeper phenotypes with the outcome measures tracked in the EHR.

Intervention(s)

NA

Outcome

The extension uses two new tables and two vocabularies to the OMOP CDM to address the structural and semantic requirements to support imaging research. The tables provide the capabilities of linking DICOM data sources as well as tracking the provenance of imaging features derived from those images.

Conclusion

Our objective is to standardize the representation of medical image events and features within the OMOP CDM framework. We introduced two new tables seamlessly integrating imaging data into the existing CDM structure, including the ability to use the OMOP cohort tools. The extended data model offers a comprehensive and unified approach for imaging research and outcome studies utilizing imaging features.

Statement of Impact

The medical imaging extension enables researchers to define computational phenotypes using imaging features. Including imaging features within the OMOP CDM broadens the scope of observational research, allowing for more comprehensive investigations into the associations between imaging biomarkers and various clinical outcomes.

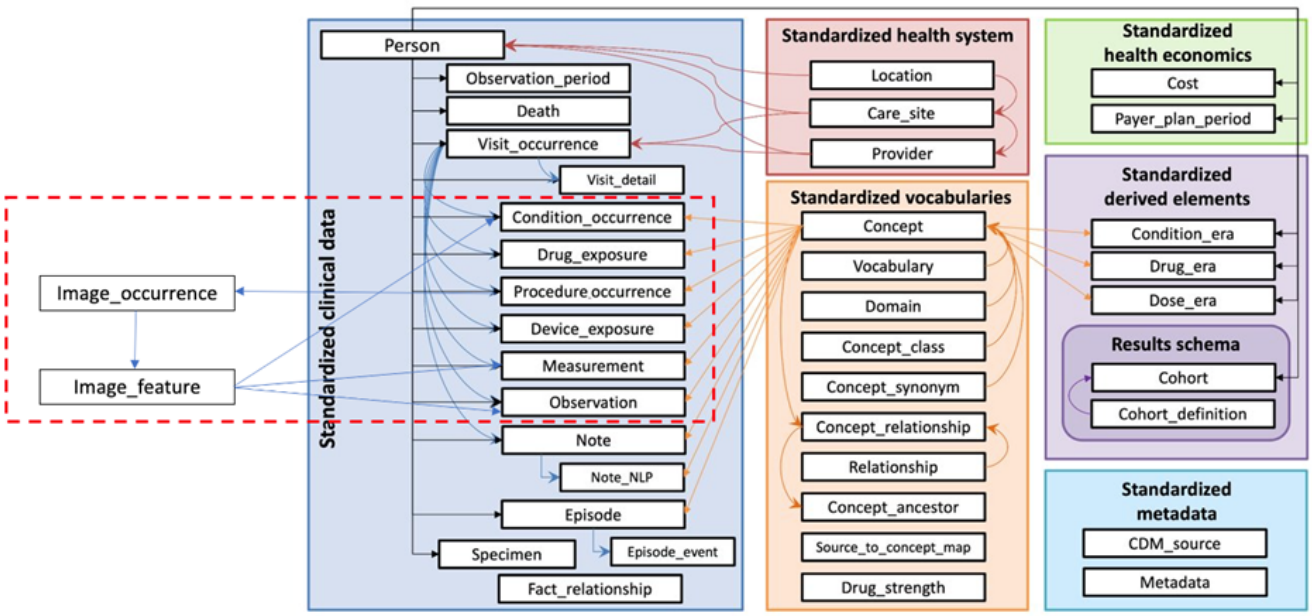


Figure 1. Incorporation of proposed Medical Image Data Model to existing OMOP CDM v5.4
HAPYGCNX-1580836-2-ANY.pdf
Data dictionary for the Image_occurrence and the Image_feature tables

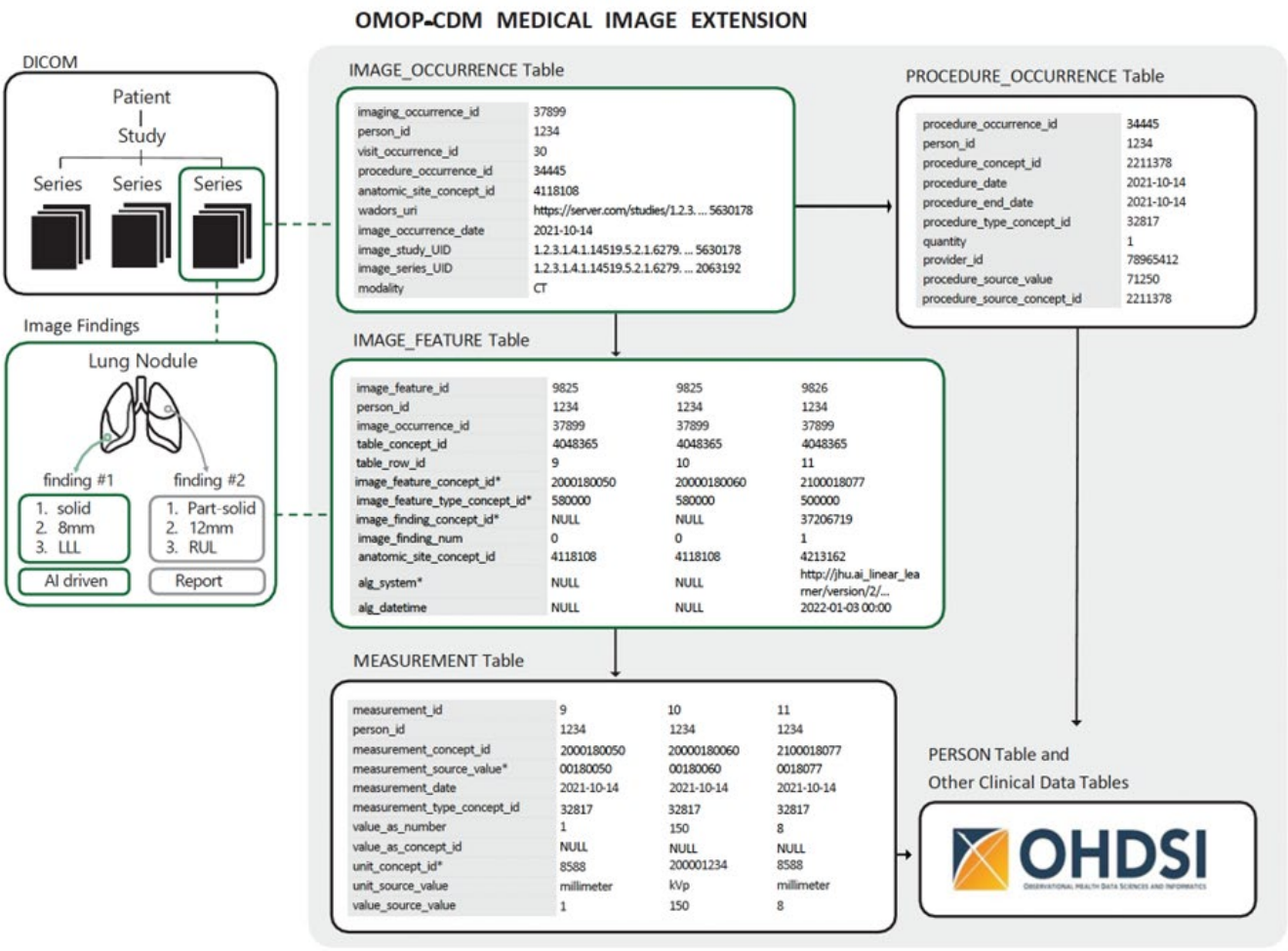
Table 2. Image_occurrence table

Field	Required	Data type	Description
image_occurrence_id (PK)	Yes	integer	The unique key is given to an imaging study record (often referred to as the accession number or imaging order number)
person_id (FK)	Yes	integer	The person_id of the Person for whom the procedure is recorded. This may be a system-generated code.
procedure_occurrence_id (FK)	Yes	integer	The unique key is given to a procedure record for a person. Link to the Procedure_occurrence table.
visit_occurrence_id (FK)	No	integer	The unique key is given to the visit record for a person. Link to the Visit_occurrence table.
anatomic_site_concept_id (FK)	No	integer	Anatomical location of the imaging procedure by the medical acquisition device (gross anatomy). It maps the ANATOMIC_SITE_SOURCE_VALUE to a Standard Concept in the Spec Anatomic Site domain. This should be coded at the lowest level of granularity.
wadors_uri	No	varchar (max)	A Web Access to DICOM Objects via Restful Web Services Uniform Resource Identifier on study level.
local_path	Yes	varchar (max)	Universal Naming Convention (UNC) path to the folder containing the image object file access via a storage block access protocol. (e.g., \\Server\Directory)
image_occurrence_date	Yes	date	The date the imaging procedure occurred.
image_study_UID	Yes	varchar (250)	DICOM Study UID
image_series_UID	Yes	varchar (250)	DICOM Series UID
modality	Yes	varchar (250)	DICOM-defined value (e.g., US, CT, MR, PT, DR, CR, NM)

Table 3. Image_feature table

Field	Required	Data type	Description
image_feature_id (PK)	Yes	integer	The unique key is given to an imaging feature.
person_id (FK)	Yes	integer	The person_id of the Person table for whom the the procedure is recorded. This may be a system-generated code.
image_occurrence_id (FK)	Yes	integer	The unique key of the Image_occurrence table.
table_concept_id	Yes	integer	The concept_id of the domain table that feature is stored in Measurement, Observation, etc. This concept should be used with the table_row_id.
table_row_id	Yes	integer	The row_id of the domain table that feature is stored.
image_feature_concept_id	Yes	integer	Concept_id of standard vocabulary—often a LOINC or RadLex of image features
image_feature_type_concept_id	Yes	integer	This field can be used to determine the provenance of the imaging features (e.g., DICOM SR, algorithms used on images)
image_finding_concept_id	No	integer	RadLex or other terms of the groupings of image feature (e.g., nodule)
image_finding_num	No	integer	Integer for linking related image features. It should not be interpreted as an order of clinical relevance.
anatomic_site_concept_id	No	integer	This is the site on the body where the feature was found. It maps the ANATOMIC_SITE_SOURCE_VALUE to a Standard Concept in the Spec Anatomic Site domain.
alg_system	No	varchar (max)	URI of the algorithm that extracted features, including version information
alg_datetime	No	datetime	The date and time of the algorithm processing.

Figure 2. Example of imaging extension tables to the OMOP clinical data tables for a lung nodule case



Keywords
Data Standardization; Observational Research; Federated Learning

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Do General Purpose Large Language Models Outperform Domain-Specific NLP Methods for Radiology Report Label Extraction?

Cody Savage, MD, Radiology Resident, University of Maryland Medical Intelligent Imaging (UM2ii) Center; Steven Rothenberg, MD; Andrew D. Smith, MD, PhD; Paul H. Yi, MD, MS

Introduction
Traditional natural language processing (NLP) methods for radiology report labeling often require fine-tuning on reports that have been manually annotated – an extremely time-consuming task. Current NLP methods are also error-prone, which introduces noise into radiology imaging datasets. Recent advances in large language models (LLMs) such as Generative Pre-trained Transformer-4 (GPT-4) have outperformed traditional methods while demonstrating domain-agnostic non-medical language task abilities. We evaluated if GPT-4 could outperform a radiology domain-specific state-of-the-art NLP tool (CheXpert) in identifying abnormalities in radiology reports.

Hypothesis
The general purpose GPT-4 model will have significantly higher accuracy for identification of abnormalities in radiology reports than CheXpert.

Methods
The radiology report “Findings” text and MeSH classification labels of 200 reports from the Indiana chest x-ray dataset were collected. The MeSH labels were used to define structured abnormal imaging finding labels, which included any disease or post-surgical imaging finding (Table 1). Reports were classified as “abnormal” if any abnormal imaging findings were present (n= 142; 71%). GPT-4 was prompted to 1) provide a list of all abnormal imaging findings present in each radiology report and 2) to classify each report as “abnormal” or “normal” without (GPT-4 base) and with advanced prompt engineering (A-GPT) comprised of ‘chain of thought’ prompting, Reflexion, and dialog-enabled resolving agents. The report text was also classified as normal or abnormal by a rules-based NLP algorithm (Stanford CheXpert labeler) to compare performance to a traditional NLP method. GPT-base report classification accuracy was compared to CheXpert using McNemar’s test.

Results
CheXpert and GPT-base correctly classified 91.5% (183/200) and 99% (198/200) of reports as abnormal or normal, respectively (P = .0003, Figure 1). A-GPT corrected 1 of the 2 misclassified reports by GPT-base (99.5% accuracy). There were 373 abnormal imaging findings in the 142 abnormal reports. Of these, GPT-base correctly identified 95.2% (355/373). A-GPT corrected 44% (8/18) of the missed abnormal findings for an accuracy of 97.3%.

Conclusion
General-purpose LLMs can extract abnormal imaging findings from report text with accuracy exceeding that of previous domain-specific NLP tools, which can be further enhanced with prompt engineering and without the need for fine-tuning.

Statement of Impact
General-purpose LLMs can serve as a universal classifier of radiology reports and extract structured abnormal imaging findings with high accuracy.

Report number	Findings	MeSH	Post-surgical findings	Reference Standard	Normal vs Abnormal	Total abnormal finding number
1	The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. Cholecystectomy clips are present. Small T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded lungs.	Osteophyte/thoracic vertebrae/multiple/small, Thickening/pleura/apex/bilateral, Lung/hyperdistention/mild	Cholecystectomy clips	Cholecystectomy clips, osteophyte, pleural thickening, hyperexpanded lungs	Abnormal	4
2	The heart size and cardiomediastinal silhouette are normal. There is hyperexpansion of the lungs with flattening of the hemidiaphragms. There is no focal airspace opacity, pleural effusion, or pneumothorax. There are multilevel degenerative changes of thoracic spine.	Lung/hyperdistention, Diaphragm/bilateral/flattened, Thoracic Vertebrae/degenerative/multiple, Emphysema	None	Hyperexpanded lungs, flattened diaphragm, degenerative changes of the spine	Abnormal	3
3	Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear. No pneumothorax or pleural effusion. No acute osseous findings.	Normal	None	None	Normal	0

Table 1 Reference standard of structured abnormal imaging finding labels. Three example reports with their corresponding MeSH terms (red) from the Indiana Chest X-ray dataset are shown. For report 1, the post-surgical finding “cholecystectomy clips” (blue) was added to the reference standard. For report 2, the MeSH term “Emphysema” was not included in the reference standard since it was not explicitly stated in the “Findings” text. Reports with no imaging findings listed in the reference were classified as “normal”.

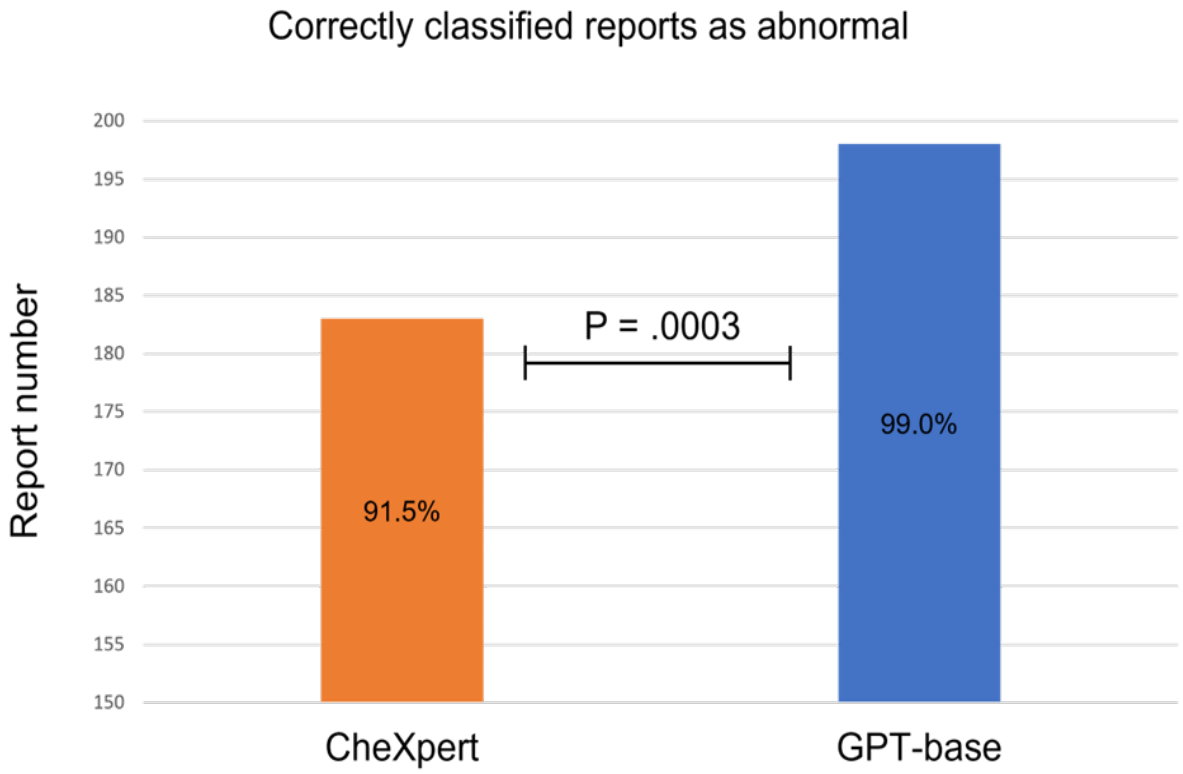


Figure 1 Classification performance comparison of CheXpert and GPT-base. The percent of the reports correctly classified as abnormal or normal by CheXpert and GPT-base are shown. GPT-base correctly classified 99% (198/200) of the reports, outperforming CheXpert at 91.5% (183/200)(P = .0003).

Keywords
Natural language processing; Text classification; Large language models; Generative transformers; Data extraction GPT

Enhancing Efficiency and Performance in Healthcare: A Federated Learning Approach for CT Image Segmentation

Alan McMillan, PhD, Professor of Clinical Health Sciences, UW Health; Iman Z. Estakhraji, PhD; John W. Garrett, PhD; Kristopher Kersten

---

Introduction

In the rapidly advancing field of healthcare, federated learning provides an efficient approach to collaborative model training across multiple sites while ensuring data privacy. This study explores the application and effectiveness of federated learning in a simulated healthcare environment.

Hypothesis

We hypothesize that using federated learning across six simulated sites would improve the performance of each site and contribute to more efficient use of computational resources. Additionally, we expect that sites with smaller and rarer datasets would equally contribute to model enhancement.

Methods

We simulated six sites with a heterogeneous distribution of 20,000 CT images and their corresponding labels for muscle segmentation. We employed Monai's SwinUNETR network for segmentation and NVidia's NVFlare package to conduct federated learning, and observed model performance across several training epochs and rounds.

Results

Our results indicated an improvement in the performance of each site, notably for the site with the smallest dataset. Sites with smaller datasets and unique or rare cases also made substantial contributions. Notably, we achieved a significant reduction in training time per epoch—from two days using single GPU training to just a few hours with federated learning.

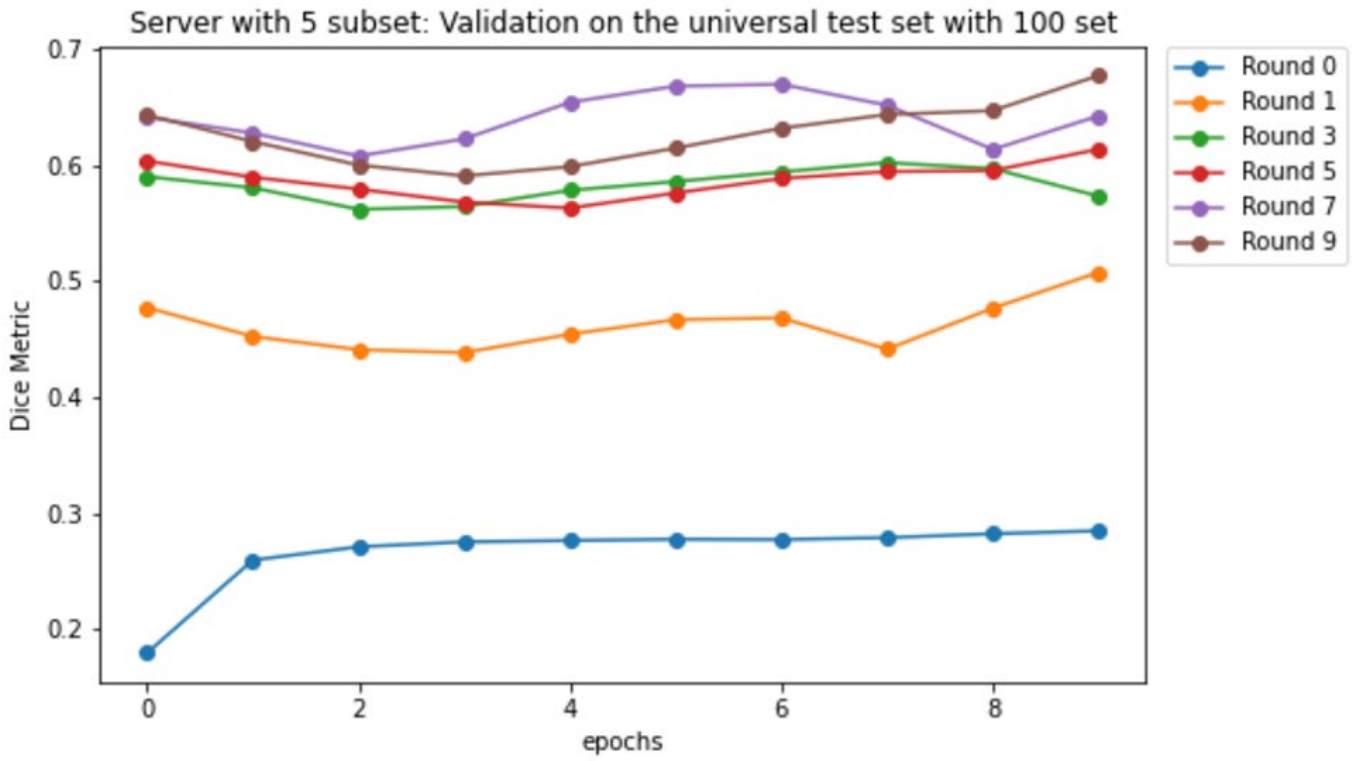
Conclusion

Federated learning significantly bolstered the performance of individual sites and contributed to more efficient use of computational resources. However, effective implementation requires careful consideration of data distribution heterogeneity, data privacy, communication overhead, and model convergence under non-IID conditions.

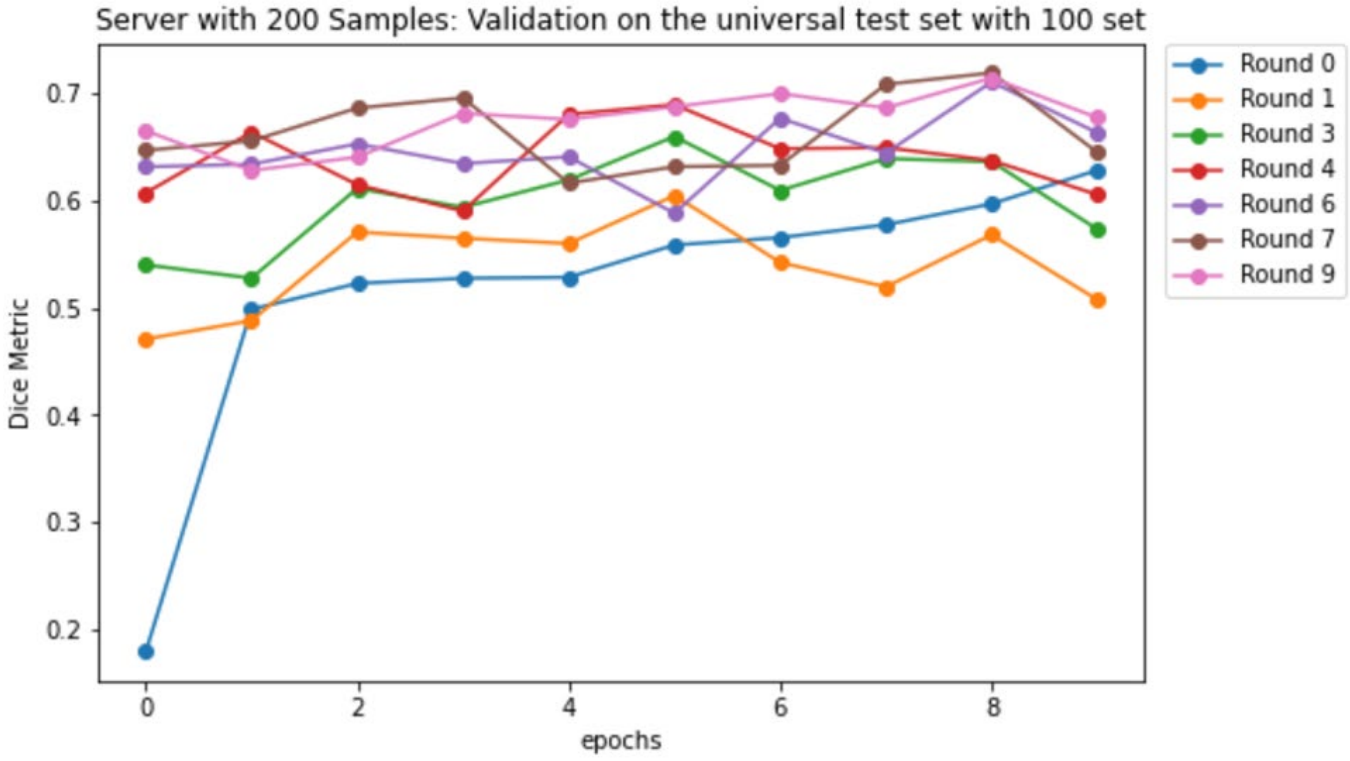
Statement of Impact

Our findings demonstrate the real-world applicability and benefits of federated learning in healthcare, particularly in resource-constrained environments. This study provides a pathway for improved and more efficient models, contributing to the advancement of collaborative healthcare research while preserving data privacy.

A site which only has 5 sets of data. It performance gets amazingly better after each round.



A site which only has 200 sets of data. It contributes to the performance of there sites.



Keywords
Federated Learning; CT Image Segmentation; NVFlare; SwinUNETR; GPU Efficiency

Enhancing Radiology Reports and Medical Information Retrieval through Question and Answering with Large Language Models

Nitin Gupta, MS, Student, University of Chicago; Ananth Prayaga, MS; Jialin Wu, MS; Hsin-Yen Yeh, MS
Utku Pamuksuz, PhD; Benan Akca, PhD

Introduction

Radiology reports are vital for diagnosis, treatment planning, and patient care, but the increasing volume of reports poses challenges for physicians. The workload intensifies as more studies are conducted, making report creation and review time-consuming, potentially leading to burnout or missed details. Our research addresses these challenges and has broader information retrieval applicability.

Hypothesis

Our value proposition is an Open Book Question and Answering (Q&A) capability on a comprehensive knowledge platform encompassing radiology and other medical reports. Utilizing open-source large language models (LLMs), the solution can provide insightful responses from the extensive information within reports, enabling thorough diagnosis, seamless report creation, and enhanced treatment.

Methods

Our solution utilizes open-source frameworks & LLMs to respond promptly and effectively to queries. Reports are initially ingested to be split into chunks and generated into embeddings. Using a Vector storage system helps with indexing and fast embeddings retrieval. Upon receiving a query, relevant context is added and passed as a combination to a chosen LLM to obtain an accurate response. Whenever an answer is not found within the context, the process proceeds to instruct/fine tune.

Results

The research, tested on over 1 million radiology reports, has yielded promising results. Our solution mitigates the risk of providing inaccurate responses and safeguards protected health information (PHI) from exposure to external third-party systems by exclusively retrieving data internally. Within a minimal response time, it can deliver precise answers to critical inquiries from healthcare professionals regarding specific patients, incorporating their comprehensive historical clinical background data stored within the system.

Conclusion

Our proposal could establish a universally adopted solution for healthcare professionals worldwide. This research encompasses various medical documents, such as electronic medical records (EMRs), pathology reports, operative reports, consultation reports, and other relevant sources. By incorporating these additional sources of information, the aim is to enhance radiologists' ability to retrieve pertinent data and finalize an imaging study effectively. This expanded application seeks to empower radiologists and physicians with a comprehensive and thorough analysis of

patients, developing a deeper understanding of their conditions and facilitating faster reporting processes.

Statement of Impact

By enhancing efficiency, resource allocation, and decision-making capabilities, our research improves patient care and alleviates healthcare professionals' workload. Optimizing AI technology utilization in the healthcare sector leads to better health outcomes and a more effective healthcare system.

High-level architecture of the proposed system

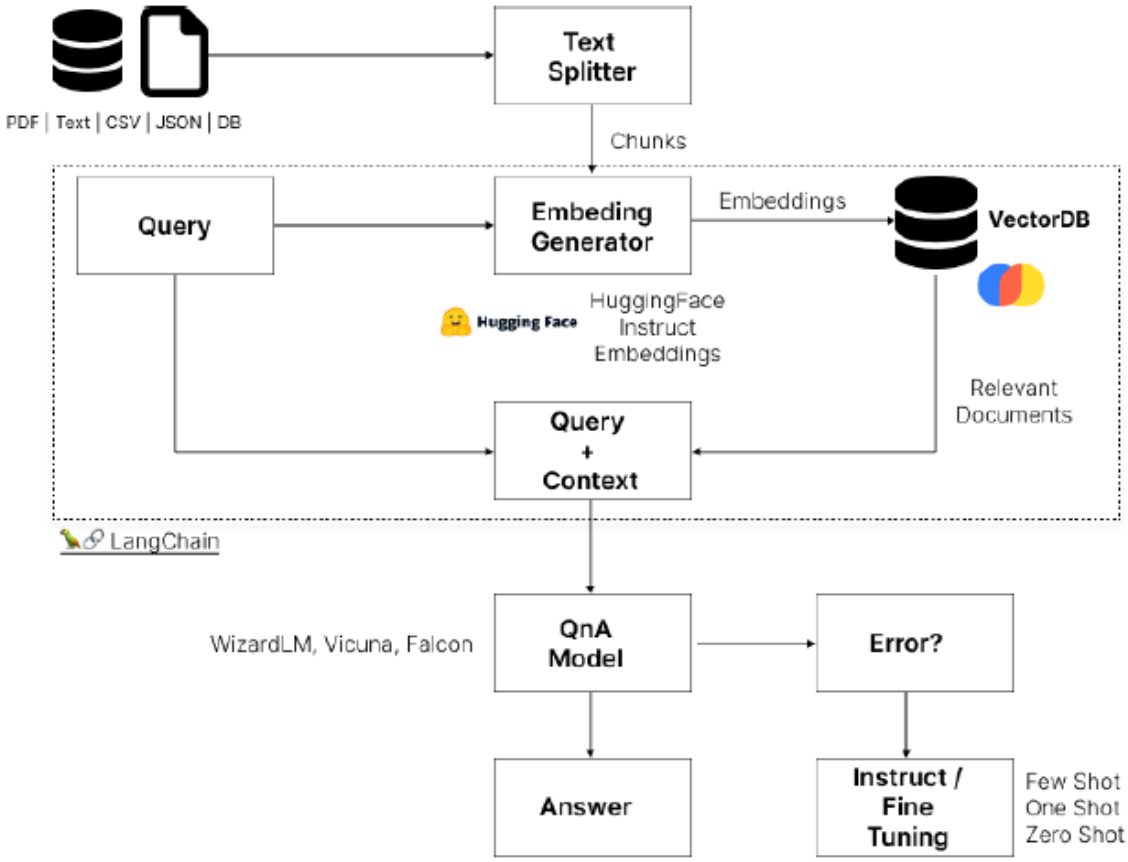


Figure: High-level architecture of the proposed system

Keywords

Artificial Intelligence; Generative AI; Large Language Models; Radiology Reports; Efficiency optimization; Open source technology

Evaluating the Diagnostic Performance of a Deep Learning Model for Detecting Thyroid Nodule Malignancy: An Expert Evaluation Study

Sanaz Vahdati, MD, Postdoctoral Research Fellow, Mayo Clinic; Bardia Khosravi, MD, MPH, MHPE; Kathryn Robinson, MD; Pouria Rouzrokh, MD, MPH, MHPE; Mana Moassefi, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction

Thyroid cancer is the most common endocrine malignancy. Ultrasound is the primary imaging modality for evaluating thyroid nodules. Based on the radiologist's assessment, further management, including fine needle aspiration, which is an invasive and costly procedure, may be requested. Thyroid nodule assessment relies on the expertise of radiologists and is subjective to an intrareader agreement. In recent years many deep-learning applications have been developed for thyroid nodule characterization. However, the evaluation of the deep learning model's performance in real-world radiology settings has been limited.

Hypothesis

We aim to evaluate the performance of our previously developed model for thyroid nodule classification into benign and malignant with a radiologist using the American College of Radiology Thyroid Imaging Reporting and Data Systems scoring(TIRADS).

Methods

The proposed model was trained on the ultrasound images of thyroid nodules from 983 patients with confirmed diagnostic reports from 2008 to 2018. 81 cases were held out as a testing set, and the rest of the data was used for training purposes. One radiologist with more than ten years of experience in ultrasound imaging evaluated the same images of the test set based on the TIRAD scores. The radiologist's thyroid nodule evaluation was recorded while the radiologist was blinded regarding the model's prediction and final pathological diagnosis of the patients.

Results

The whole pipeline from the developed model reached an AUROC of 0.84 (CI 95%: 0.75-0.91) with sensitivity and specificity of 84% and 63%, respectively. The TIRAD evaluation of the test set had a sensitivity of 76% and specificity of 34% which was comparably lower than the model's prediction (p-value=0.003). A comparison of our model's performance with ground truth and the TIRAD score based on an expert radiologist's evaluation was analyzed. In 28% of the cases, the model predicted benign nodules as benign with TIRADS >3 reports from an expert radiologist. In 19% of cases, our model predicted benign nodules with TIRADS ≤3 as benign, and in 33% of cases, our model predicted malignant nodules correctly with the TIRADS >3. In addition, our model predicted no nodule as malignant, while the nodule had a benign biopsy report with a TIRADS ≤3.

Conclusion

We demonstrate the potential diagnostic performance of the deep learning model by comparison of its predictions with TIRAD scores from an expert radiologist.

Statement of Impact

Thyroid nodule assessment in ultrasound is subjective to inter and intra-reader agreement. Our deep learning model can provide further assistance to radiologists for thyroid nodule evaluation.

Table 1. Comparing the model's performance with the ground truth and ACR-TIRAD score.

Ground truth	Prediction	TIRAD score	Percent
Benign	Benign	TIRAD>3	28.4 (23 of 81)
Benign	Benign	TIRAD≤3	18.5 (15 of 81)
Benign	Malignant	TIRAD≤3	0
Benign	Malignant	TIRAD>3	6.17 (5 of 81)
Malignant	Malignant	TIRAD>3	33.3 (27 of 81)
Malignant	Malignant	TIRAD≤3	3.7 (3 of 81)
Malignant	Benign	TIRAD>3	6.17 (5 of 81)
Malignant	Benign	TIRAD≤3	3.7 (3 of 81)

Keywords

Deep learning; TIRADS; Thyroid nodule; Ultrasound

Extraction of Labels from Radiology Reports using ChatGPT

Jason Adleberg, MD, Radiology Resident, Mount Sinai; Nicholas Primiano, MD; Alexander Kagen, MD; Tessa S. Cook, MD, PhD, CIIP, FSIIM

Introduction

Deep learning models in radiology require large amounts of labeled data. Creation of such datasets can be time-consuming and error-prone. Natural Language Processing (NLP) models, such as ChatGPT, have recently gained popularity due to their ability to comprehend unstructured text. In this project, ChatGPT is used to extract labels from chest radiograph reports, and compared to human performance.

Hypothesis

ChatGPT cannot extract labels from radiology reports with comparable accuracy to human performance.

Methods

200 chest radiographs and their corresponding reports were pulled from the MIMIC database, an open-source collection from Beth Israel Deaconness Hospital in Boston, MA. A radiology resident reviewed each of these 200 reports, and marked the presence or absence of three findings: (1) cardiomegaly, (2) pleural effusion, and (3) pneumothorax. A fourth category for (4) normal, or negative chest radiograph was also created. This annotation was done manually to create a ‘gold-standard’ set. ChatGPT was then asked to do the same task on 200 exams. The human and ChatGPT outputs were compared, using metrics of precision, recall, F1 score, and accuracy.

Results

For 200 reports, macro-averaged values for precision, recall, F1 score, and accuracy were 0.827, 0.932, 0.851, and 0.940, respectively. For each individual condition, accuracy was 0.925 for cardiomegaly, 0.965 for pleural effusion, 0.995 for pneumothorax, and 0.875 for a normal chest radiograph. Incorrect classifications were mostly due to medical terminology, such as misclassification of “hydropneumothorax” and “blunting of the costophrenic angle” as “no pleural effusion”. Other misclassifications were due to language, such as “borderline cardiomegaly’. Full details and examples are shown in the included tables.

Conclusion

ChatGPT can extract labels from radiology reports with comparable accuracy to human performance. In the future, NLP applications can be used to create large labeled datasets for the development of artificial intelligence applications.

Statement of Impact

ChatGPT can extract labels from radiology reports with comparable accuracy to human performance. In the future, NLP applications can be used to create large labeled datasets for the development of artificial intelligence

applications.

Correct and Incorrect Classification Numbers for the Four Labels.

Finding	*True Positive*	*True Negative*	*False Positive*	*False Negative*
Cardiomegaly	49	136	2	13
Pleural Effusion	48	145	4	3
Pneumothorax	52	147	1	0
Normal CXR	20	155	25	0

Correct and Incorrect Classification Examples for the Four Labels.

Finding	*True Positive*	*True Negative*	*False Positive*	*False Negative*
Cardiomegaly	"Borderline cardiomegaly is stable" "Enlargement of the cardiac silhouette"	"Heart size top normal"	"Top normal cardiac silhouette"	"Enlargement of the cardiac silhouette"
Pleural Effusion	"Small left pleural effusion"	"No pleural effusion"	"Pulmonary edema"	"Blunting of costophrenic angle" "Hydro-pneumothorax"
Pneumothorax	"Apical pnemothorax" "Hydro-pneumothorax"	"No pneumothorax"	"Air within the right mediastinum"	N/A
Normal CXR	"No acute cardiopulmonary process"	"Bilateral consolidations"	"Resolution of pneumothorax, chest tube in place" "Unchanged left upper lobe mass"	N/A

Keywords
ChatGPT; NLP; AI

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

From the Operating Room to the Office: Digital Staining of White Light Cystoscopy Videos for Bladder Cancer Detection

Greyson Wintergerst, Undergraduate Researcher, Vanderbilt University; Shuang Chang, PhD; Haoli Yin; Kristen R. Scarpato, MD, MPH; Amy N. Luckenbaugh, MD; Sam S. Chang, MD, MBA; Soheil Kolouri, PhD; Audrey K. Bowden, PhD

Background/Problem to be solved

Bladder cancer is the tenth most common malignancy and also carries the highest treatment cost of all cancers. Much of this cost comes from the need for repeated cystoscopy due to its high recurrence rate. White Light Cystoscopy (WLC) is used to examine the bladder for suspicious lesions, but suffers from limited sensitivity. Blue Light Cystoscopy (BLC) utilizes a fluorescent dye to stimulate contrast, which improves sensitivity by 43%. However, the cost of the BLC system and the lengthy dye administration procedure preclude its use in the office. Here, we propose the first use of digital staining, or image-to image translation (I2I), to convert WLC images into accurate BLC-like images, thereby increasing the accessibility of BLC for the office without cost.

Intervention(s)

Data Collection/Pre-Processing: In total, 45 videos containing region-matched WLC and BLC frames were collected from 31 patients using a clinical BLC system. We first innovated a color adaptation workflow to account for the variations in brightness, color, and saturation across clinical videos, minimizing out-of-distribution data. Optimal transport was then applied to move the RGB color point cloud of each frame to a reference color distribution, determined by minimizing the Wasserstein distances of the associated WLC or BLC frames, shown in Figure 1. Architecture and Assessment: Our model was trained to transform color-normalized WLC frames into digitally generated BLC (dgBLC) images on unpaired WLC/BLC images following the Density Changing Regularized Unpaired Image Translation (DECENT) model, (Xie, NeurIPS,2022) as shown in Figure 2. The collected WLC and BLC frames were not perfectly registered. While not required to perform the transformation on test data, manual registration of near-perfectly matched WLC and BLC pairs was performed using fitgeotform2d in MATLAB to permit assessment with clinically relevant metrics. Evaluation was performed with a fluorescence segmentation mask we developed for assessing digital-staining accuracy.

Outcome

Figure 3 shows example testing data. These preliminary results demonstrate successful generation of digitally stained images. Fluorescent labeling of tumor regions in the generated images is shown for various morphological appearances.

Conclusion

In this study, we have addressed the heterogeneity in clinical data distributions to enable successful digital staining of WLC/generated BLC-like images with an unpaired I2I model. We also developed a novel, clinically relevant method to assess the digital staining performance and demonstrated successful staining.

Statement of Impact

We present the first digital staining of clinical WLC frames into BLC frames. This study paves the way for a cost-effective alternative to BLC for in-office examinations.

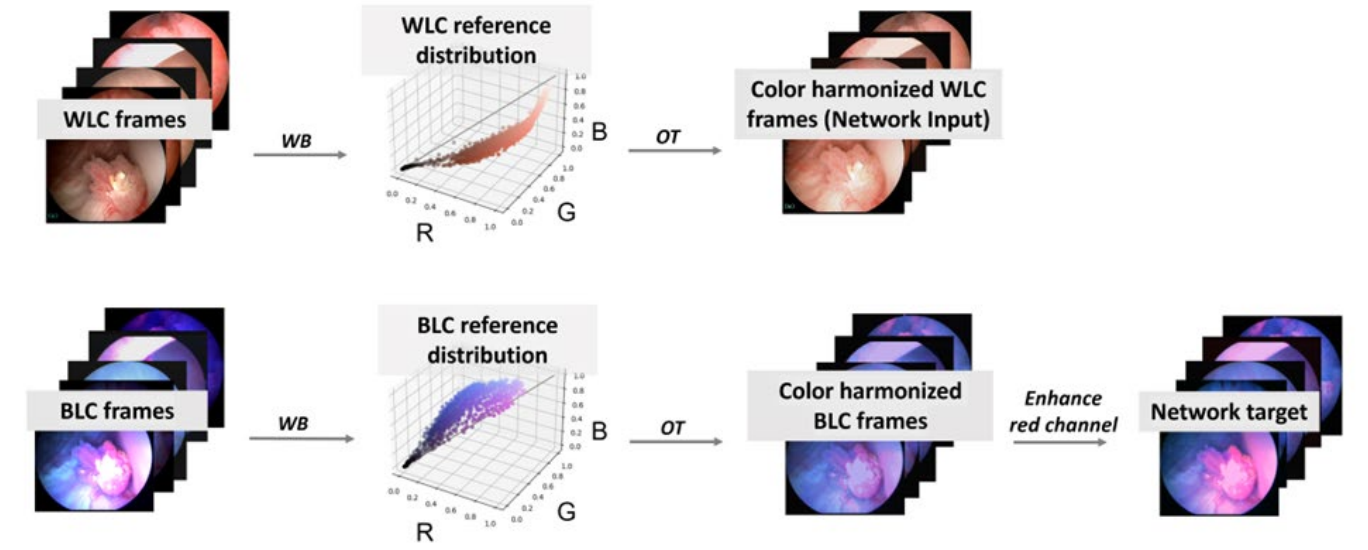


Figure 1: Color harmonization and optimization process for WLC and BLC frames to align color distributions. To normalize the color for each frame, we use optimal transport (OT) to transform the color distribution of each input frame to match a reference distribution based on the Wasserstein barycenter (WB) calculated from a subset of the WLC/BLC frames. This process yields BLC images with muted fluorescence, so the red channel is enhanced using the original BLC images as a reference.

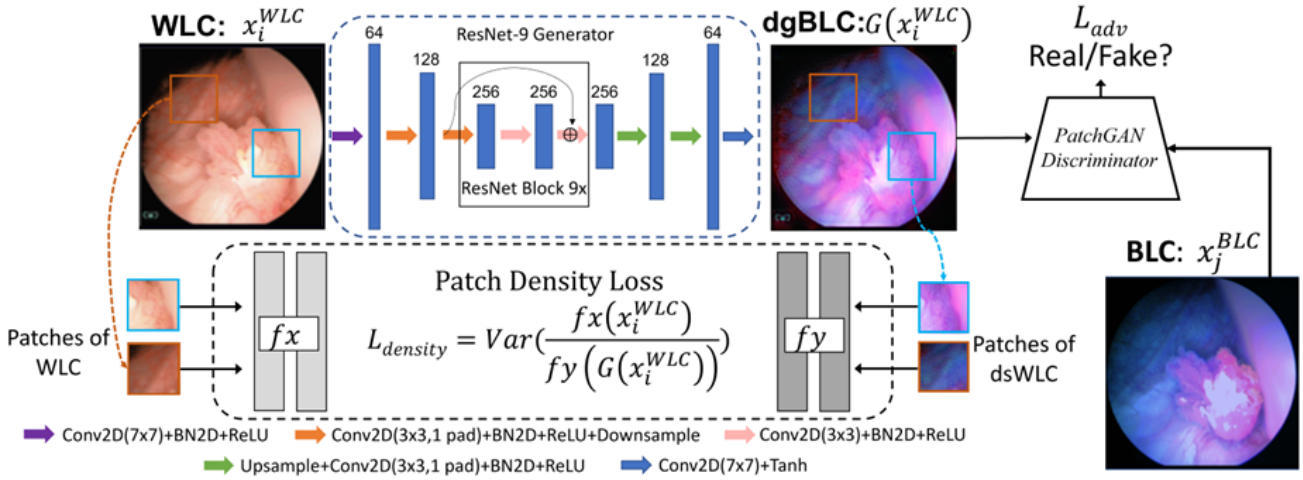


Figure 2: Patch-density-based GAN model used to perform style-transfer I2I translation. The blue-dashed box represents the generator that produces dgBLC images from WLC inputs. We employed autoregressive flows for density estimation (represented by fx and fy) and used a ResNet-based generator with a PatchGAN discriminator. The model consists of three terms from the original method: an adversarial loss, an identity mapping loss, and a density-changing loss. The network outputs are the digitally stained WLC frames, or dgBLC frames.

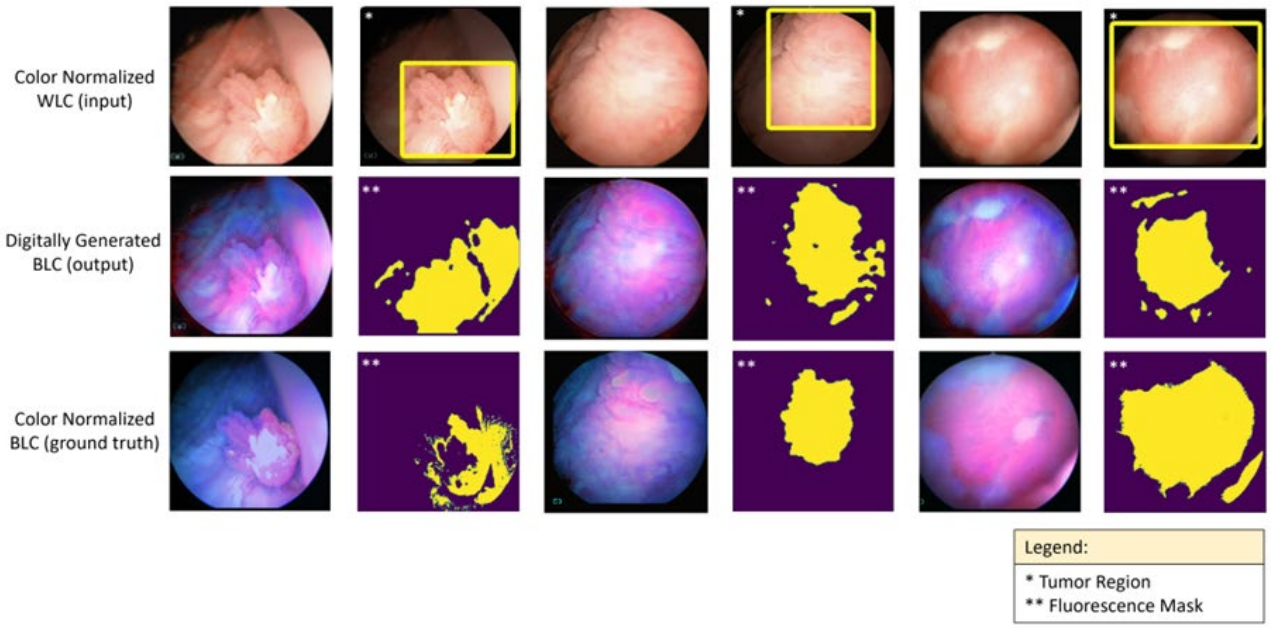


Figure 3: Preliminary results depicting examples of digitally generated BLC frames from the testing data, along with the accompanying fluorescence mask. According to the legend, the top WLC images with yellow-boxed regions (*) are the expected tumor regions (i.e., where fluorescence is expected). Images with two asterisks (**) depict the fluorescence masks applied on the BLC images, utilizing the YCbCr color space and color channel intensity thresholds. These allow for quantitative, clinically relevant comparison between generated vs. ground truth frames.

Keywords
Cystoscopy; I2I Translation; Registration; Normalization; Bladder cancer; GAN

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Generation of Radiology Report's Impression from Findings' Description on Pediatric Abdomen Ultrasound

Dana Alkhulaifat, MD, Postdoctoral Research Fellow; Children's Hospital of Philadelphia; Vahid Khalkhali, PhD; Patricia P. Rafful, MD, PhD; Michael Welsh, PhD; Susan T. Sotardi, MD

Introduction

Training and validation of artificial intelligence (AI) models require annotated data. Annotation is challenging due to the need for well-defined search criteria (e.g., inclusion and exclusion), domain-specific knowledge and expertise, budget, and time. Additionally, since establishing well-defined annotation rules poses numerous challenges, manually curated datasets can be subjective and inconsistent. Natural language processing has demonstrated utility in the identification of diagnoses from radiology reports. These reports could also be used to annotate images, thereby improving the speed and accuracy of data curation for model building. In this research, we trained an AI model to generate an impression from the findings section on pediatric ultrasound reports.

Hypothesis

We examined whether a generative deep learning (DL) model can construct an accurate impression from findings similar to a radiologist in pediatric ultrasound reports.

Methods

After receiving IRB exemption, a total of 16891 reports of pediatric complete abdominal ultrasound were obtained from our institution (51% female, 47% male, 2% unknown), from 2015 to 2022 (age range (0, 43 years), mean of 7y8m, std of 6y4m). The findings and impression sections of each report were extracted. The dataset was divided into (60%, 20%, 20%) for the training, validation, and testing of a generative DL model, respectively. Recall-Oriented Understudy for Gisting Evaluation (Rouge) 1, 2, L, and Lsum were used as performance metrics. Three outcomes were established to measure the interpretability performance: “human impression is more complete”, “human and machine impressions are similar”, and “model impression is more complete.” These outcomes were assigned to 300 random reports from the test dataset and results were computed and declared with 95% confidence level.

Results

The Rouge 1, 2, L, and Lsum metrics were (0.61, 0.49, 0.57, 0.6) (0.52, 0.38, 0.48, 0.5) on the train and validation datasets, and (0.51, 0.37, 0.46, 0.49) on the blind test dataset. The average completeness scores of the generated impressions were (36.6%, 59.7%, 3.7%) which indicates that 63.4% of generated impressions are similar or better than human impressions with confidence interval of (0.5857, 0.6943). Table 1 demonstrates an example of a radiologist’s impression and an AI-generated impression.

Conclusion

The Rouge and human scores indicate that a generative DL model fine-tuned on the reports of pediatric complete abdominal ultrasounds can be a valuable and time-saving tool for annotating datasets for the purpose of building AI models in radiology.

Statement of Impact

The utilization of radiology reports to annotate images improves the speed and accuracy of large dataset curation for model building.

Table 1: Examples of radiologists' impressions and AI-generated impressions for the same report.

Example	Radiologist's Impression	AI-Generated Impression
1	No adrenal hemorrhage or intra-abdominal hemorrhage. Small amount of simple free fluid in the abdomen.	1. Decreased peristalsis of the visualized bowel segments. 2. Gallbladder sludge without evidence of cholecystitis. 3. Mild increased echogenicity of the bilateral kidneys, likely physiologic in this neonate. 4. Mild hepatomegaly.
2	Large amount of alimentary debris in the stomach. No definite abdominal mass.	Large amount of food within the stomach. Otherwise, normal examination of the abdomen.
3	No discrete focal lesion within the liver, spleen, or kidneys. No sonographic evidence of acute intra-abdominal process.	Normal abdominal ultrasound.

Table 1: Examples of radiologist's impressions and AI generated impressions for the same report.

Keywords

Report Generation; Pediatric Ultrasound Abdomen Complete; Generative Deep Learning

Large Language Model Improves Detection of Negated Expressions in Radiology Reports

Yonatan Babore, Medical Student, University of Pennsylvania; Yvonne Su; Charles E. Kahn, MD, MS, FACR, FSIIM

Introduction
Named-entity recognition (NER) and negation detection are crucial to extract structured information from unstructured text. The extracted information can provide meaningful insights for clinical decision-making, patient care, and medical research. Large Language Models (LLMs), such as the Bidirectional Encoder Representations from Transformers (BERT), may improve understanding of text.

Hypothesis
This study compared the ability of a rule-based NLP system and an LLM—both trained for the medical domain—to correctly detect negated concepts in radiology reports.

Methods
This HIPAA-compliant, IRB-approved study used a convenience sample of 1000 consecutive de-identified radiology reports from a large U.S.-based academic health system. Duplicate reports were excluded to yield the study cohort of 984 reports. The medspaCy system (Eyre et al., AMIA Annu Symp Proc 2021) was applied to identify terms from the vocabularies of the Unified Medical Language System (UMLS), the Radiology Gamuts Ontology, and RadLex in the report cohort. An LLM, the Clinical Assertion and Negation Classification BERT (CAN-BERT; van Aken et al., NLPMC 2021) had been trained using MIMIC III and patient data from the i2b2/VA challenge. We compared medspaCy and CAN-BERT to detect negations of the identified entities. Power analysis determined a sample size of 59 pairs to achieve $\alpha = 0.05$, $\beta = 0.8$ for McNemar's test. 200 entities were randomly selected and manually labeled by two independent researchers to create the ground truth. Outputs of the two models were compared with ground-truth annotations using McNemar's test. Precision, recall, and F1 were computed.

Results
In total, 63,797 entities were extracted from the reports and subjected to negation annotation by each model. For negation detection, medspaCy attained recall of 0.806, precision of 0.382, and F1 of 0.518. CAN-BERT achieved recall of 0.962, precision of 0.807, and F1 of 0.877; it significantly outperformed the medspaCy model ($p < 0.001$).

Conclusion
CAN-BERT detected the negation of entities in radiology reports with high precision and recall. This LLM’s performance significantly exceeded that of a rule-based NLP tool for negation detection.

Statement of Impact
The findings of this study could facilitate informed decision-making when choosing an NLP tool for processing radiology reports, guiding the development of more efficient and reliable NLP systems.

	medspaCy **CORRECT**	medspaCy **INCORRECT**
BERT **CORRECT**	154	38
BERT **INCORRECT**	3	5
McNemar's $\chi 2 = 28.195$, p-value $= 1.097 \times 10^{-7}$		

Table 1. Comparison of performance in each model with ground truths

Keywords
negation expression detection; radiology reports; natural language processing; large language model; transformer model

Identifying Cerebrospinal Fluid Leak using Brain MRI: A Deep Learning Approach

Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic; Mana Moassefi, MD; Ian T. Mark, MD; Ajay Madhavan, MD; Jared T. Verdoorn, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; John C. Benson, MD

Introduction

Cerebrospinal fluid (CSF) leaks cause an increase in disability and mortality rates if left untreated. Accurate diagnosis prevents complications such as meningitis, brain abscesses, and headaches. Imaging tests such as contrast-enhanced T1 (CET1) MRI are used to diagnose CSF leaks. The documented features of the MRI with leakage include sagging of the brain, enhancement of the pachymeninges, engorgement of venous structures, pituitary enlargement, and subdural fluid collections. A deep learning (DL) model might provide a more accurate detection of leakage-related features and other shape changes of the venous sinuses. Our team has created a completely automated classifier that distinguishes between patients with CSF leakage, those without, and individuals whose MRI features did not clearly diagnose CSF leakage. Our model can detect CSF leakage-related features while eliminating the need for subjective visual assessments and time-consuming distance measurements.

Hypothesis

DL models can be utilized in the diagnosis of CSF leak using brain MRI.

Methods

We acquired sagittal CET1 images of 151 patients, who were categorized as having either functional CSF leak, no CSF leak, or indeterminate based on their imaging findings by our institutional neuroradiologists. To ensure a balanced distribution of labels, we divided our dataset into 5folds at the patient level stratified by labels. To evaluate the reliability of our model, we conducted 5-fold cross-validation. Using a 3D Densenet121, we trained the model to classify the three aforementioned classes. Our evaluation was based on the area under the receiver operating characteristic (AUC) for each validation fold we reported.

Results

The classifier demonstrated an average AUC of 0.86, with a standard deviation of 0.07 across the folds, in determining the positive, negative, and indeterminate cases for CSF leak.

Conclusion

Based on brain MRI, we developed a DL model that can predict the presence of the CSF leak. However, further model refinement and external validation are necessary before clinical adoption. Developing a fast, automated, and reliable model for detecting CSF leaks using brain MRI enhances diagnostic accuracy and streamlines clinical workflow.

Statement of Impact

This research highlights the significant potential of deep learning (DL) in diagnosing cerebrospinal fluid (CSF) leaks through the use of brain MRI.

Keywords

Deep learning; CSF leak; Classification

Towards Trustworthy Deep Learning: Applying Mondrian Conformal Prediction to Intracranial Hemorrhage Detection

Cooper Gamble, Undergraduate Research Intern, Mayo Clinic; Shahriar Faghani, MD

Introduction
Deep learning (DL) has earned its place in medical imaging by performing classification, segmentation, generation, and detection tasks with impressive efficiency and accuracy. By default, out-of-the-box DL tools give statistical estimates about the certainty of their predictions, but these figures cannot be interpreted as probabilities or confidence scores because they are uncalibrated. Our approach replaces these estimates with statistical guarantees for predictions by calibrating DL outputs. We apply Mondrian Conformal Prediction (MCP) to Intracranial hemorrhage (ICH) detection, one of many severe conditions for which DL models have been trained as clinical support systems.

Hypothesis
Mondrian Conformal Prediction supports trustworthy deep learning by providing statistical guarantees for model predictions.

Methods
We selected the CQ500 dataset, which is composed of 193,317 scans from 491 patients. It contains challenging cases (n=52539) in which one or more of the readers disagree about a patient's diagnosis, and we designated the remaining patients' slices as definite cases. We trained a YOLOv8 model on both positive-only and balanced subsets of our dataset to localize and identify the ICH types. Finally, we performed MCP by sorting the confidence scores of our calibration set. We reported MCP's accuracy in flagging challenging cases.

Results
The mAP of 0.547 during validation and 0.411 during testing on definite samples. Our model was best at localizing and identifying IVH, IPH, and SDH instances with respective mAPs of 0.995, 0.694, 0.443 during validation. At a p-value threshold of 0.2, MCP identified 100% of challenging cases by generating a prediction set which contained both the presence and absence of at least one hemorrhage type.

Conclusion
We built an ICH detection model whose performance rivals state-of-the-art models, but which can also output prediction sets with statistical guarantees. Furthermore, applying MCP to improve trustworthiness demonstrated perfect accuracy in flagging challenging cases. Continued external validation is necessary for clinical adoption, but this study is a promising first step in that direction.

Statement of Impact
Trust is a critical component of practical deep learning tools. We offer a deployable, statistically rigorous, and task- and model- agnostic approach to increase trustworthiness in DL models by calibrating and filtering their predictions.

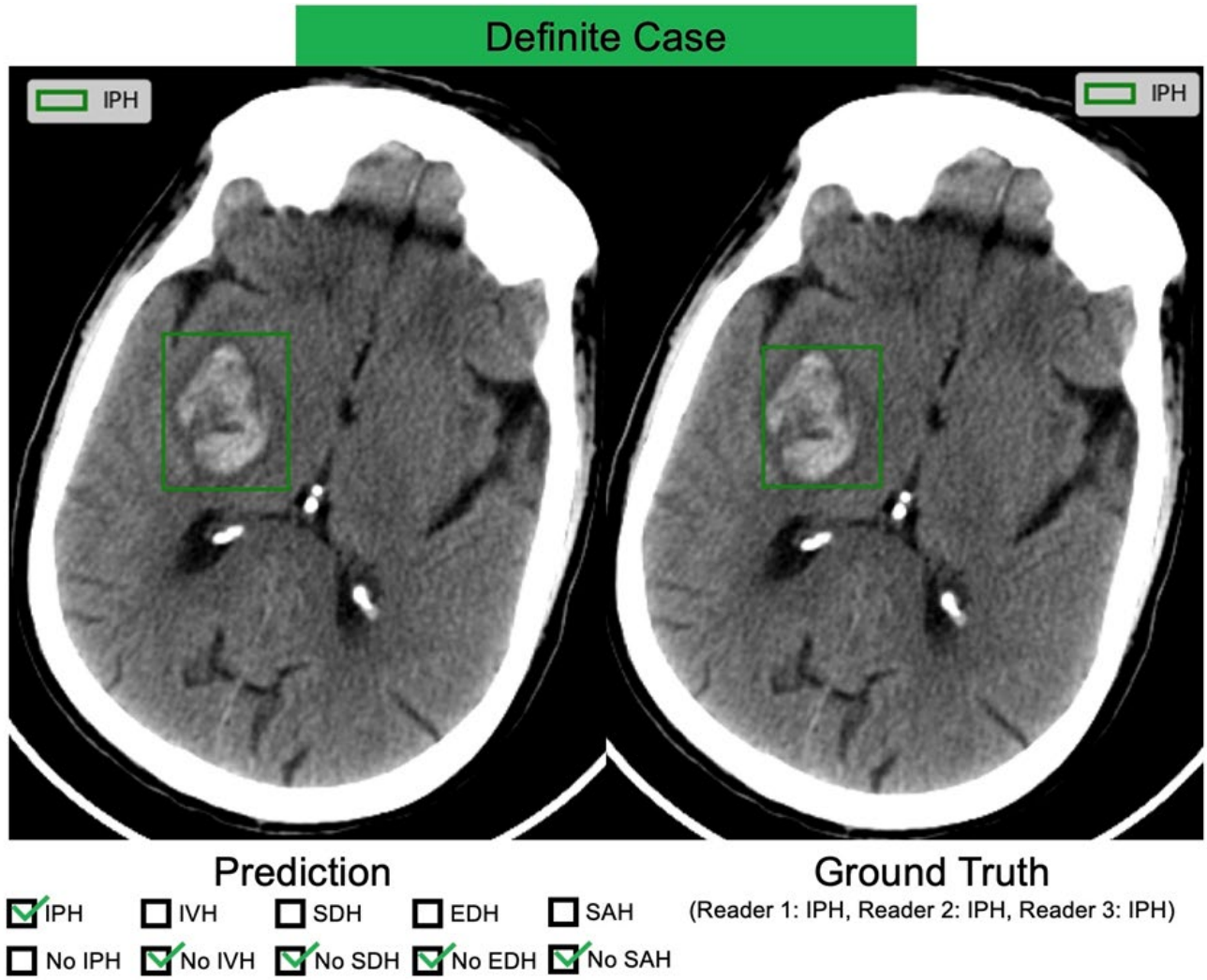


Figure 1: Prediction and ground truth for a definite case. Checked boxes indicate labels that were included in the model's prediction set.

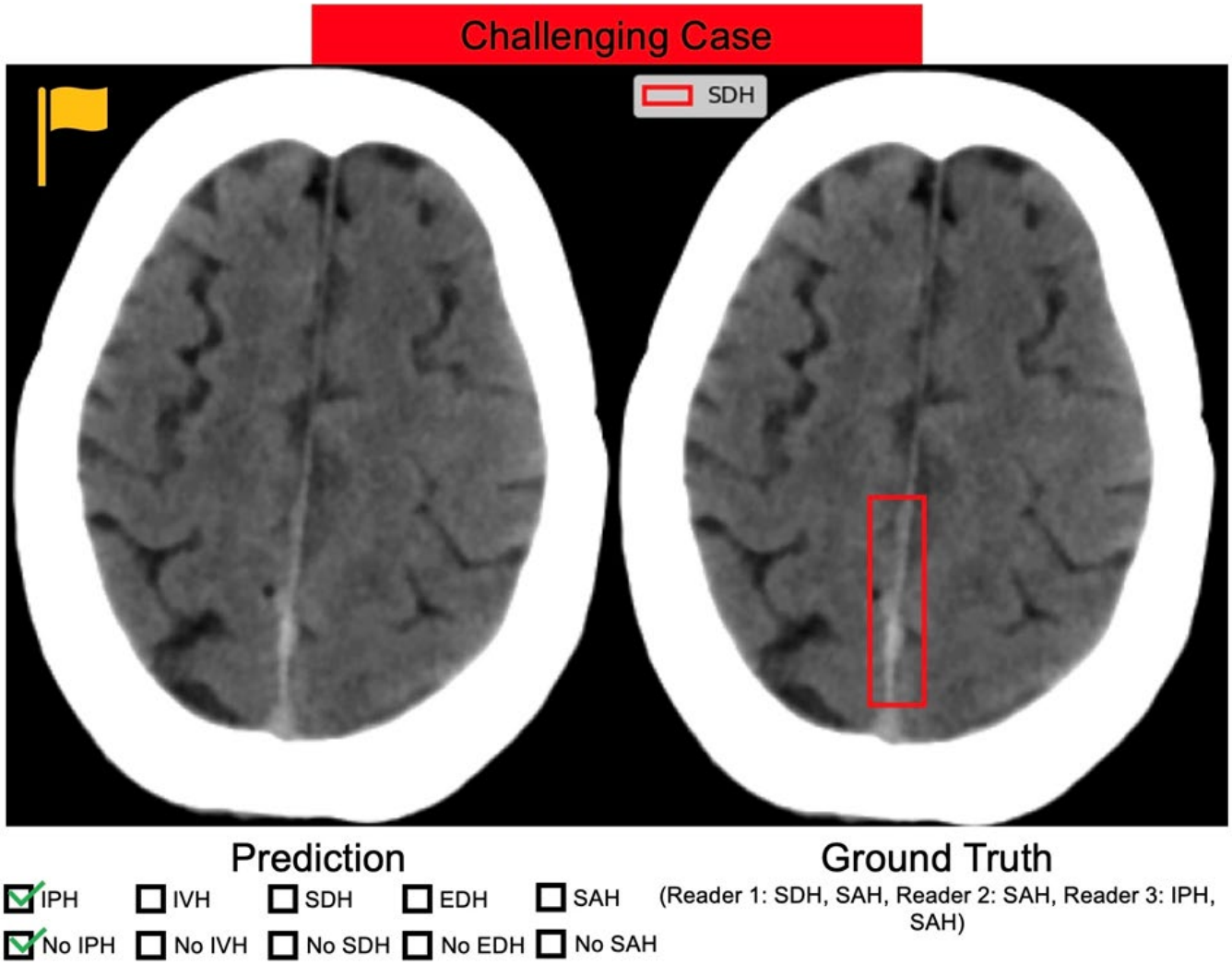


Figure 1: Prediction and ground truth for a challenging case. Checked boxes indicate labels that were included in the model's prediction set.

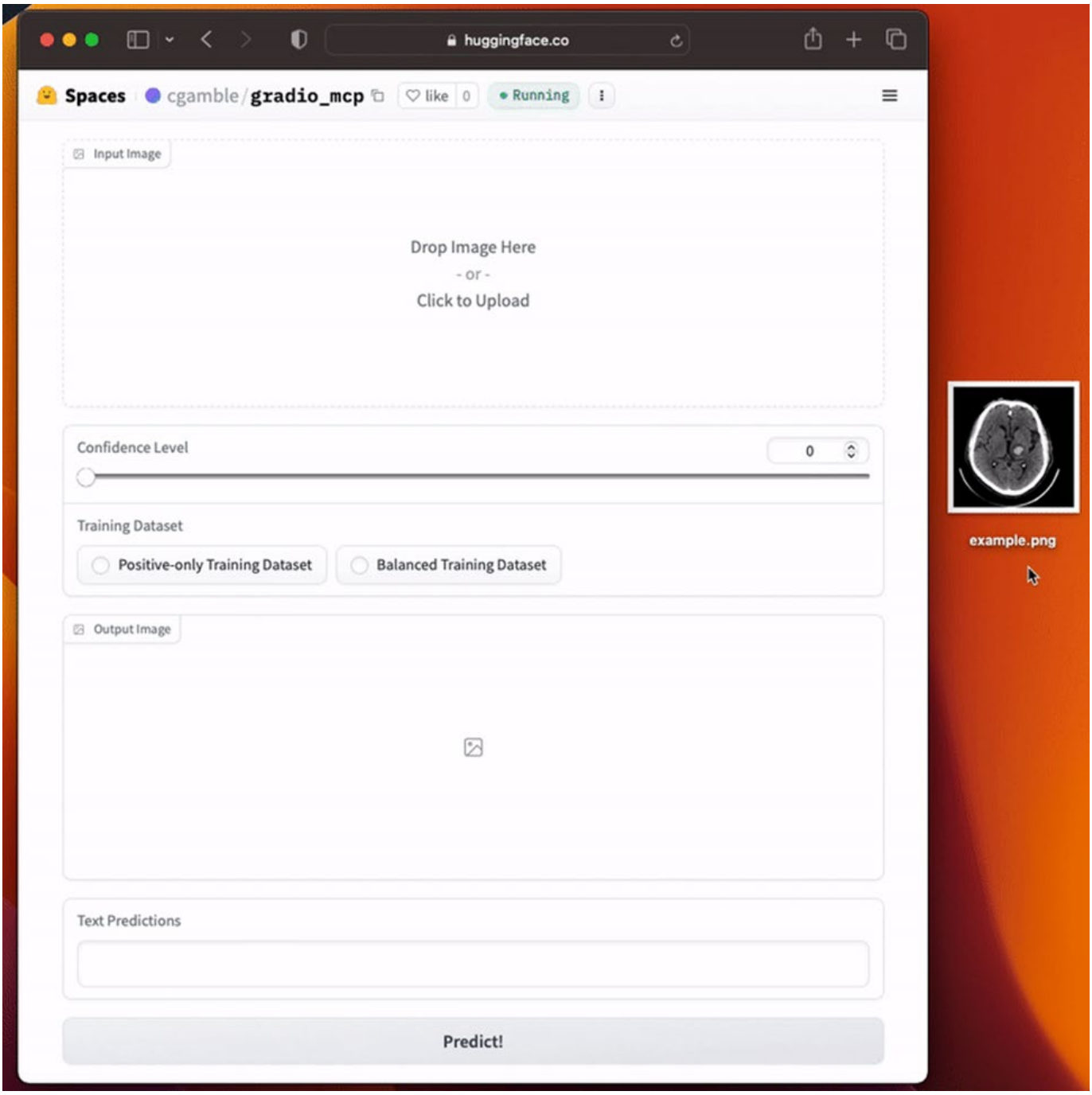


Figure 3: Interface for Mondrian Conformal Prediction with YOLOv8 model (available at https://huggingface.co/spaces/cgamble/gradio_mcp).

Demo link: https://www.abstractscorecard.com/uploads/Tasks/upload/20687/HAPYGCNX-1583345-3-ANY(3).gif

Keywords
Conformal Prediction; Brain Hemorrhage; Uncertainty Quantification; Object Detection

Uncertainty Quantification in Radiogenomics Analysis Using Mondrian Conformal Prediction

Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic; Mana Moassefi, MD; Khosravi Bardia, MD, MPH, MHPE; Gian Marco Conte, MD, PhD; Pouria Rouzrokh, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction

Recent research has demonstrated the potential of deep learning (DL) models in predicting genetic status based on medical imaging. Nevertheless, the extent to which DL can accurately predict the status of all genes remains uncertain. Notably, a growing body of literature suggests that the methylation status of the O-6-methylguanine-DNA methyltransferase (MGMT) gene promoter in glioma cannot be determined using MRI. To address this issue, the present study investigates the application of Mondrian Conformal Prediction (MCP) in the classification of isocitrate dehydrogenase (IDH) and MGMT in glioblastomas utilizing MRI. MCP assigns conformity scores to predictions, allowing the identification of certain and uncertain cases. By focusing on certain cases and disregarding the uncertain ones, it is anticipated that the model's performance will improve if the previously observed low performance was due to uncertainty rather than the model's ability to distinguish between classes.

Hypothesis

MCP improves radiogenomics classification performance metrics by filtering out uncertain cases.

Methods

The publicly available The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset, which comprises MRI sequences and corresponding tumor genetic profiles, was utilized in this analysis. The dataset consisted of MGMT negative: positive samples in a ratio of 1:3.34 and IDH wildtype: mutant samples in a ratio of 1:3.81. The dataset was divided into 5 folds, stratified by labels at the patient level. Two 3D-Densenet121 models were trained on the T2 sequence for predicting IDH and MGMT. The efficacy of MCP in IDH and MGMT classification was evaluated.

Results

The accuracy and AUC for IDH classification were 0.83 and 0.88, respectively, while for MGMT prediction, they were 0.78 and 0.54. With the application of MCP, the accuracy for IDH prediction improved to 1, but no improvements were observed in MGMT classification.

Conclusion

The findings emphasize that MCP enhances model performance when the model can distinguish between different classes and further support the notion that current classifiers struggle to effectively capture MGMT signals from MRI data.

Statement of Impact

By filtering out uncertain cases, Mondrian Conformal Prediction (MCP) enhances the performance of deep learning models, particularly when there is a sufficient signal for the task at hand, resulting in improved accuracy and other performance metrics.

Keywords

Deep learning; Uncertainty quantification; Conformal prediction; Glioma; IDH; MGMT

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Leveraging 3D Segmentation Datasets for Rapid Body Region Classification

Xue Li, MS, Research Assistant, University of Wisconsin-Madison; Nicolas Pannetier, PhD; Mehul Sampat, PhD; Travis Richardson; Richard Bruce, MD; John W. Garrett, PhD; Alan B. McMillan, PhD

Introduction

The growing utilization of deep learning into medical imaging workflows warrants the development of efficient identification of the input body region for quality assurance and downstream algorithm selection. However, the development of these models requires significant resources for data curation if developed exclusively for body region classification. Comprehensive whole-body segmentation datasets, although laborious and time-consuming to compile, offer immense potential in intricate anatomy detection, providing automated diagnostics, and facilitating radiotherapy or surgical planning. However, practical constraints arise from the use of 3D segmentation models in high throughput workflows due to substantial computational and resource demands compared to 2D classifiers. This research aims to establish that high-quality 3D whole-body segmentation datasets can be conveniently adapted to train more efficient 2D body region classification algorithms.

Hypothesis

Whole-body segmentation datasets can be readily adapted into 2D body region classification models.

Methods

We utilized the TotalSegmentator dataset (https://totalsegmentator.com/) which contains CT images from 1204 unique subjects, captured via various scanners and protocols, with 104 distinct anatomical structures annotated. Each 3D CT image was flattened to 2D in the coronal plane and normalized min-max. Training, validation, and test splits were 60%, 20%, and 20% respectively. A 3-layer U-Net model with 1.7 million parameters using 64x64 input patches was trained using MONAI for 800 epochs.

Results

The model was evaluated on the test dataset with ~1.4 seconds running time per subject using < 600 MB RAM on a commodity GPU (NVIDIA 1080Ti). Evaluation metrics included accuracy, precision, recall, and F1 score. Micro-level evaluation yielded an accuracy of 0.8204, precision of 0.8344, recall of 0.8689, and an F1 score of 0.8513. Similarly, on a macro-level, the model achieved an accuracy of 0.8132, precision of 0.8403, recall of 0.8585, and an F1 score of 0.8367. The 2D model offers robust predictive performance with a good balance between precision and recall, as evidenced by the high F1 scores. An example subject is shown in Figure 1.

Conclusion

We demonstrate the feasibility of employing 3D segmentation datasets to train efficient 2D classification models for expedited body region classification in CT images. Whereas a segmentation algorithm could require 100x computation time, the 2D model enabled rapid analysis with substantially lower resources, potentially enabling real-time evaluation on low resource hardware.

Statement of Impact

Whole-body segmentation datasets can be transformed to train efficient 2D body region classification models, enabling high throughput capability on low resource hardware, which could enhance quality assurance and downstream algorithm selection in medical imaging workflows.

Keywords

Body Region Classification; Whole-Body Segmentation; Deep Learning

mRMR-permute: Permutation Testing for Unbiased Minimum Redundancy Maximum Relevance Feature Selection

Winston T. Chu, PhD, Data Scientist, Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health; Marcelo A. Castro, PhD; Venkatesh Mani, PhD; Jens H. Kuhn, M.D., PhD; Ian Crozier, MD; Claudia Calcagno, PhD; Jeffrey Solomon, PhD

Background/Problem to be solved

Radiomics applies advanced analytic approaches to resolve and quantify the shapes, intensities, and patterns in medical images. However, radiomic feature extraction produces hundreds of features, many of which are redundant and should be removed. The minimum redundancy maximum relevance (mRMR) algorithm has emergent value for feature ranking. However, it does not provide a standardized way to threshold the features for fully automated, unbiased, and model-independent feature selection. Permutation testing, a statistical technique that enables hypothesis testing with minimal assumptions about data distribution, may be used to apply statistical thresholds to mRMR feature scores.

Intervention(s)

In this study, we present mRMR-permute, a novel method that combines mRMR and permutation testing for unbiased and model-independent feature selection. Using mRMR-permute, a standard statistical threshold (e.g., p-value < 0.05) can be applied to threshold mRMR feature scores, reducing potential user bias in feature set size selection and without model-dependence (e.g., recursive feature elimination). Here we describe the approach in detail (Figure 1) and demonstrate its applications on a public dataset (Figure 2).

Outcome

To evaluate performance, mRMR-permute was used for radiomic feature selection on a publicly available dataset generated by the TCGA Research Network (http://cancergenome.nih.gov/) and downloaded from Kaggle (https://www.kaggle.com/competitions/glioma-radiomics/data/). These data were curated to determine whether deletion of gene 1p19q in low-grade brain glioma could be predicted by radiomic features extracted from magnetic resonance imaging (MRI) scans. mRMR-permute performance was compared to other popular feature-selection techniques. When averaging across the examined model architectures, mRMR-permute outperformed least absolute shrinkage and selection operator (LASSO) regression, Elastic Net regression, an f-test p-value threshold of 0.05, and no feature selection (Figure 2A). A receiver operating characteristic (ROC) curve was generated for the best model and feature-selection method (mRMR-permute and random forest); the mean ROC area under the curve was found to be 0.75 (Figure 2B).

Conclusion

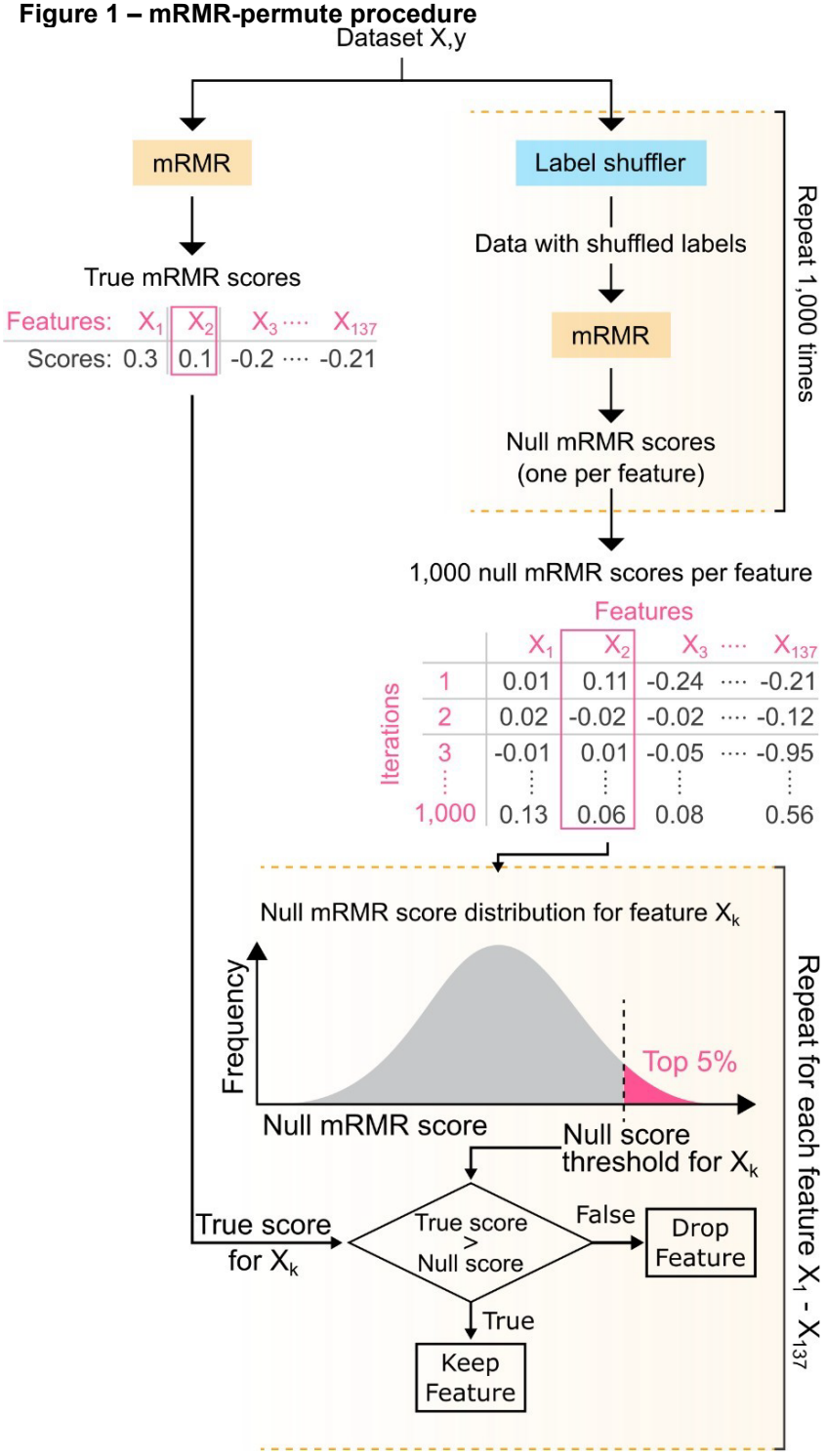
We show that mRMR-permute is an effective method for unbiased and model-independent feature selection. In a

public dataset, we demonstrated that mRMR-permute, on average, outperforms several popular feature-selection techniques across multiple model architectures.

Statement of Impact

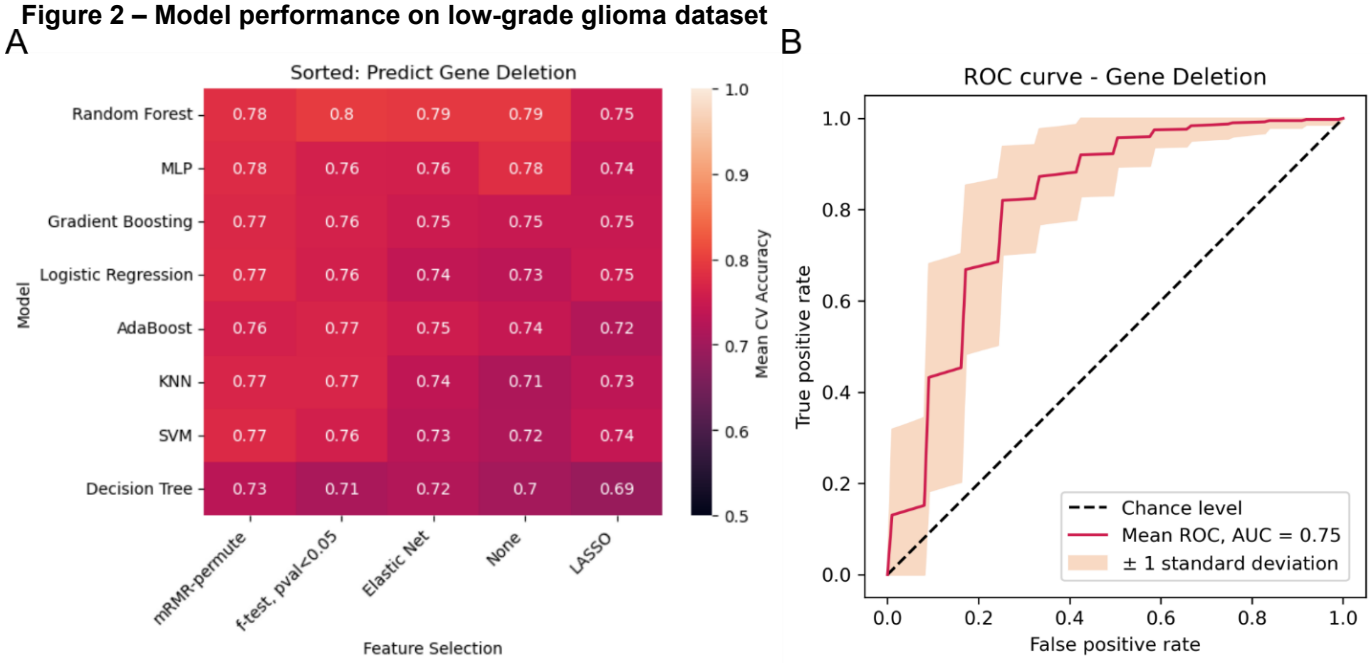
mRMR-permute may be used to effectively and automatically select radiomics features, enhancing the performance of classification models that use radiomic data. The broader impact of mRMR-permute is in its potential applications to any classical machine learning classification task.

Figure 1 – mRMR-permute procedure



The mRMR-permute procedure is as follows: 1) Shuffle the labels in the dataset. 2) Run mRMR on the shuffled dataset. 3) Record the null mRMR scores for each feature. 4) Repeat steps one through three 1,000 times. 5) For each feature, calculate the mRMR score threshold that would result in only 5% (for a *p*-value < 0.05 threshold) of null mRMR scores greater than that threshold and therefore being detected as significant. 6) Run mRMR on the original dataset (true labels). 7) For each feature, if the true mRMR score is greater than its threshold from step five, select the feature. mRMR = minimum redundancy maximum relevance

Figure 2 – Model performance on low-grade glioma dataset



A) All combinations of model architectures (y axis) and feature selection techniques (x axis) were trained in a cross-validated (CV) manner. The mean accuracy across 20 CV iterations was used to compare combinations. The rows (models) and columns (feature selection) are ordered by average performance, increasing from bottom to top and right to left such that the top-left-most cell represents the best (on average) model and feature selection technique. B) The receiver operating characteristic (ROC) curve was plotted for the top model and feature selection technique (mRMR-permute and random forest). The average curve across 20 CV iterations is shown in red and the standard deviation is highlighted in orange. AUC = area under the curve, MLP = multilayer perceptron, KNN = k-nearest neighbors, SVM = support vector machine, LASSO = least absolute shrinkage and selection operator

Keywords
feature selection; machine learning; radiomics; mRMR

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

One Copy Is All You Need: Resource-Efficient Streaming of Medical Imaging Data at Scale

Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine;
Adway Kanhere, MS; Eliot L. Siegel, MD, FSIIM; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Introduction

Large-scale medical imaging datasets have accelerated the development of artificial intelligence tools for clinical decision-making. However, the large size of these datasets is a major bottleneck for users with limited storage or bandwidth. Many users may not even require such large datasets, as they are often downsampled to lower-resolutions for deep learning [Figure 1a]. If these datasets could be downloaded directly at their desired resolution, the storage and bandwidth savings would be enormous. However, it is impossible to anticipate every users' requirements or store data at multiple resolutions. What if we could store images at a single resolution but send them at different resolutions?

Hypothesis

We developed the Medical Image Streaming Toolkit (MIST) as an open-source framework on the premise that progressive resolution can tackle infrastructural inefficiencies by reducing storage and data transfer requirements, while making data more accessible to users in resource-limited areas, without affecting downstream applications.

Methods

MIST utilizes HTJ2K, a state-of-the-art progressive codec, and provides the following capabilities: 1) encoding/decoding of medical imaging data acquired across different imaging formats and modalities, and 2) stream optimization for rapid access to sub-resolution images by splicing and streaming only a partial bytestream, sufficient to reconstruct the image at the user's desired resolution [Figure 1b]. We simulated an imaging database, hosting three large-scale datasets across various formats and modalities [Figure 2, Table 1], and calculated the storage and bandwidths savings of transmitting all three datasets to a hypothetical user using MIST. Furthermore, we quantitatively evaluated the fidelity of images streamed by MIST using structural similarity index measure (SSIM) and mean absolute deviation (MAD).

Results

Compared to conventional imaging databases, MIST significantly reduced the host's storage requirements from 194.10GB to 79.35GB (59.12%), while reducing the amount of data transmitted to the user from 194.10GB to 12.74GB (93.44%) [Table 1]. Not only were the progressively encoded images identical to the original images (SSIM=1.00, MAD=0.00), but the sub-resolution images streamed by MIST were strikingly close to images downsampled at the user's requested resolution using conventional approaches (SSIM=0.97, MAD=0.01).

Conclusion

MIST can dramatically reduce imaging data infrastructural inefficiencies by efficiently encoding, streaming, and decoding medical images, while maintaining diagnostic quality. Streaming medical images may help facilitate widespread adoption of the cloud for medical image storage and transmission.

Statement of Impact

As cloud-based imaging infrastructures gain popularity, MIST presents an open-source, format-agnostic, resource-efficient framework to dramatically reduce inefficiencies that currently limit widespread hosting and streaming of medical images over the Internet.

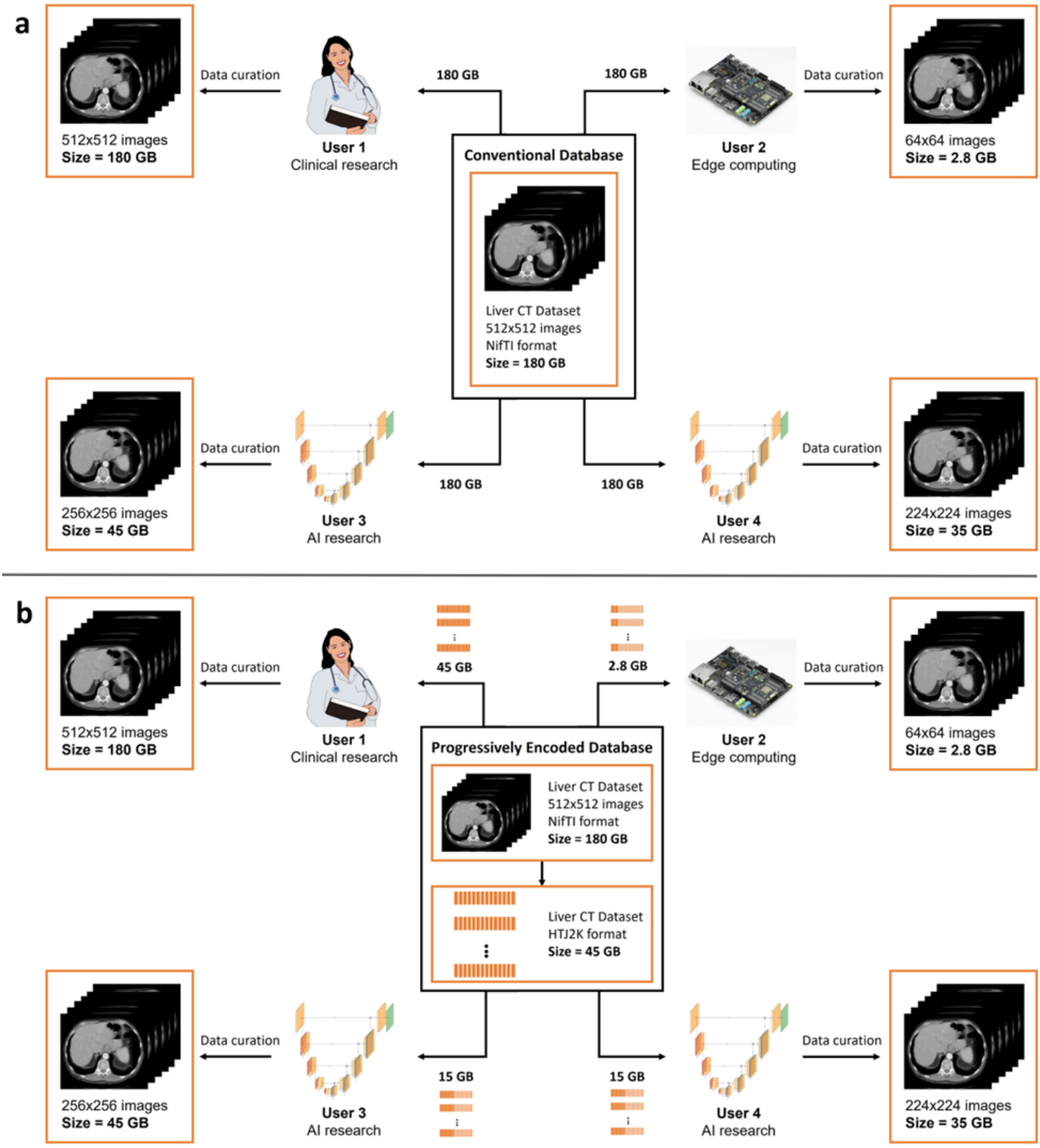


Figure 1. (a) Illustration of the conventional approach to medical imaging data curation. In this scenario, four users download a cloud-hosted liver CT dataset for different use-cases. Each use case requires a different sized image, yet the user must download the data at the baseline resolution before processing them into their desired resolutions for further analysis, resulting in inefficient use of bandwidth, storage, and compute. (b) How can progressive encoding address the computational, network and storage inefficiencies associated with current computation infrastructures developed to host medical imaging datasets? Progressively encoding the liver CT dataset produces

a storage saving of 4x for the host organization while reducing the total data transferred from the host organization by 9x. Similarly, savings in the data storage requirements for the users range from 4x–64x.

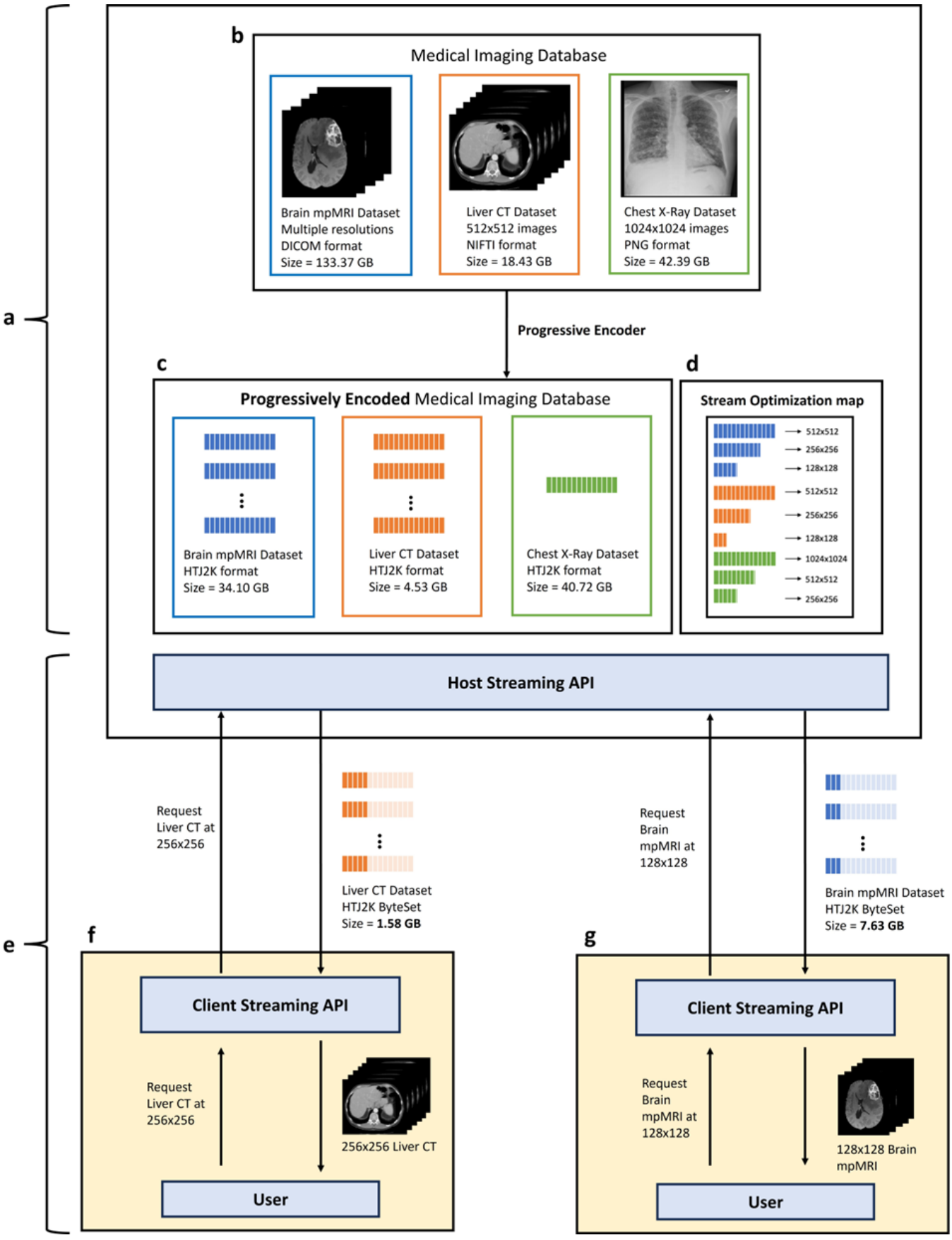


Figure 2. Illustration of the Medical Image Streaming Toolkit (MIST). (a) The progressive encoder-decoder framework encodes the medical images (b) into progressively encoded HTJ2K images (c) and generates a stream optimization map (d) between different byte subsets and image resolutions. (e) The progressive streaming framework enables the users (f-g) to request images at their desired resolutions by determining the appropriate partial bytestream and decoding them into desired imaging formats for further analysis. Using MIST, the left user (f) was able to download the entire liver data in 1.58 GB (as opposed to 18.43 GB) and the right user (g) was able to download the entire brain mpMRI data in 7.63 GB (as opposed to 133.73 GB) resulting in 15x bandwidth and storage savings.

Dataset	NIH Chest X-Ray	MSD Liver	UPENN-GBM
Modality	X-ray	CT	mpMRI
Format	PNG	NifTI	DICOM
Compression	Lossless	Lossless	-
Storage Size (GB)			
Original	42.39	18.34	133.37
MIST	40.72 (-3.95%)	4.53 (-75.28%)	34.10 (-74.43%)
Quantitative Metrics			
SSIM	1.000 ± 0.0000	0.999 ± 0.0002	1.000 ± 0.0001
MAD	0.000 ± 0.0000	0.001 ± 0.0001	0.001 ± 0.0002
Requested Resolution	224x224	256x256	128x128
Data Transmitted (GB)			
Original	42.39	18.34	133.37
MIST	3.53 (-91.68%)	1.58 (-91.38%)	7.63 (-94.28 %)
Quantitative Metrics			
SSIM	0.971 ± 0.0087	0.959 ± 0.0150	0.968 ± 0.0590
MAD	0.011 ± 0.0022	0.010 ± 0.0038	0.011 ± 0.0137

Table 1. Summary of the quantitative evaluation of MIST's encoding, decoding, and streaming efficiency, compared to the conventional approach. For the host, MIST's encoding efficiency and lossless decoding efficiency is measured. For the client, MIST's streaming and lossy decoding efficiency is measured at the user's requested resolution. Quantitative metrics reported as Mean ± SD.

Keywords
Image compression; Data efficiency; Infrastructure inefficiencies; Streaming; Resource efficiency; Progressive resolution

Text2Cohort: Democratizing the NCI Imaging Data Commons with Natural Language Cohort Discovery

Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine;
Adway Kanhere, MS; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Introduction

The NCI Imaging Data Commons (IDC) is a cloud-based data commons that provides researchers with open access to large-scale cancer imaging datasets. However, querying the IDC database for cohort discovery and access to imaging data requires technical know-how, such as executing SQL queries. We developed and tested Text2Cohort, a large language model (LLM)-based tool to automatically facilitate user-friendly and intuitive natural language cohort discovery in the IDC.

Hypothesis

Text2Cohort will allow for automated IDC cohort discovery using natural language queries without any knowledge of coding.

Methods

We developed Text2Cohort as a pipeline using GPT-3.5 – the state-of-the-art language model that also powers ChatGPT – with four major components: prompt engineering, BigQuery generation, BigQuery autocorrection, and cohort extraction [Figure 1]. Text2Cohorts translates user input into IDC database queries using prompt engineering and autocorrection and returns the query's response to the user [Figure 2]. Autocorrection resolves errors in queries by passing the errors back to the model for interpretation and correction [Figure 3]. We evaluated Text2Cohort on 50 natural language user inputs ranging from information extraction to cohort discovery. The resulting queries and outputs were verified by two computer scientists to measure Text2Cohort's accuracy and F1 score. For user inputs that generated incorrect queries and responses, the query was corrected by an expert and the Levenshtein distance, i.e., minimum number of character-edits to correct the query, was calculated.

Results

Text2Cohort successfully generated queries and their responses with an 88% accuracy and F1 score of 0.94. However, it failed to generate queries for 6/50 (12%) user inputs due to syntax and semantic errors. These six incorrect queries were manually corrected by an expert with 12.83±5.81 character-edits. In other words, our results demonstrate that despite failing to correct 10% of all queries due to semantic errors, the natural language toolkit was able to generate queries with correct structure and autocorrect errors within them with a 98% success rate.

Conclusion

Text2Cohort succeeded at generating queries with correct responses, but occasionally failed due to a lack of understanding of the data schema. Despite these shortcomings, Text2Cohort demonstrates the utility of LLMs to enable researchers to discover and curate cohorts using data hosted on IDC with high levels of accuracy using natural language in a more intuitive and user-friendly way.

Statement of Impact

Natural language cohort discovery demonstrates the utility of LLMs to democratize access to imaging data hosted by groups such as the IDC in a more intuitive and user-friendly way.

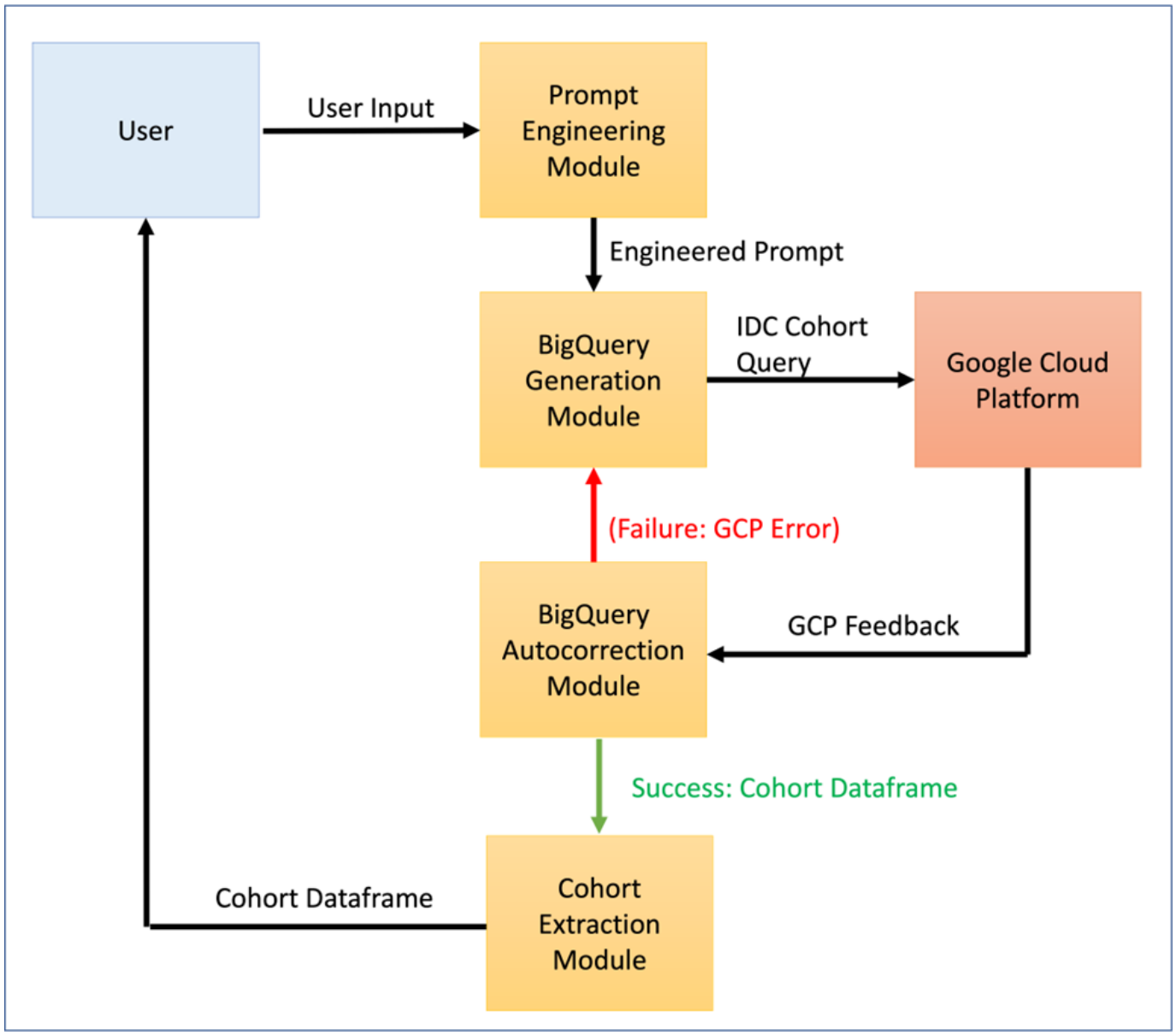


Figure 1. Illustration of the natural language cohort discovery toolkit.

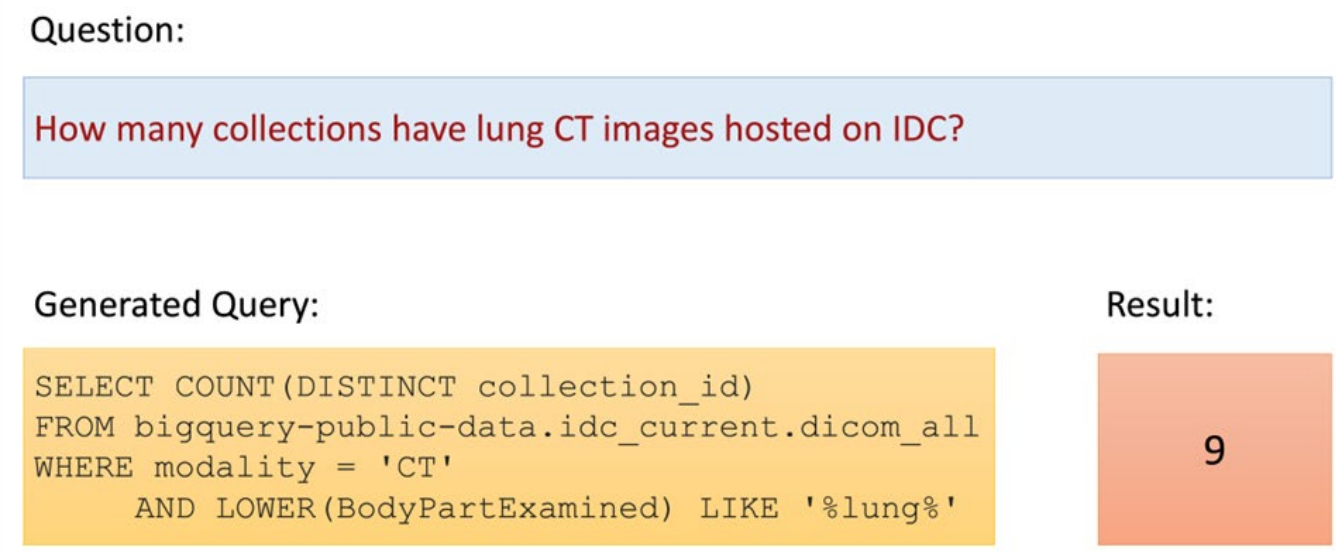


Figure 2. The natural language cohort discovery toolkit on an example natural language user input. The toolkit first transforms the user input into a query, uses the generated query to query the BigQuery table, and returns the response back to the user.

Question:

For the upenn_gbm collection, download only T2 and FLAIR scans

Generated Query:

```
SELECT * FROM `bigquery-public-data.idc_current.dicom_all`
WHERE collection_id = 'upenn_gbm' AND (scan_type LIKE '%T2%' OR scan_type LIKE '%FLAIR%')
```

GCP:

```
400 Unrecognized name: scan_type at [4:40] Location: US Job ID: a7e76e88-ac94-440f-b9f0-676879f4c3bb
```

```
SELECT * FROM `bigquery-public-data.idc_current.dicom_all`
WHERE collection_id = 'upenn_gbm' AND (REGEXP_CONTAINS(scan_protocol_name, 'T2') OR REGEXP_CONTAINS(scan_protocol_name, 'FLAIR'))
```

```
400 Unrecognized name: scan_protocol_name at [4:56] Location: US Job ID: 80c0c1ce-4b04-4884-883d-8de41f919764
```

```
SELECT * FROM `bigquery-public-data.idc_current.dicom_all`
WHERE collection_id = 'upenn_gbm' AND (REGEXP_CONTAINS(series_description, 'T2') OR REGEXP_CONTAINS(series_description, 'FLAIR'))
```

```
400 Unrecognized name: series_description; Did you mean SeriesDescription? at [4:56] Location: US Job ID: 4b54531b-40db-468f-844b-49cb2171d22f
```

```
SELECT * FROM `bigquery-public-data.idc_current.dicom_all`
WHERE collection_id = 'upenn_gbm' AND (REGEXP_CONTAINS(SeriesDescription, 'T2') OR REGEXP_CONTAINS(SeriesDescription, 'FLAIR'))
```

	collection_id	BodyPartExamined	Modality	SeriesDescription
0	upenn_gbm	HEAD_NECK	MR	AX T2 FLAIR : Processed_CaPTk
1	upenn_gbm	HEAD_NECK	MR	AXIAL FLAIR : Processed_CaPTk
2	upenn_gbm	HEAD_NECK	MR	AX T2 : Processed_CaPTk
3	upenn_gbm	HEAD_NECK	MR	AX FLAIR FS : Processed_CaPTk
4	upenn_gbm	BRAIN	MR	AXIAL T2 FLAIR: Processed_CaPTk
...	...	...	...	...
60172	upenn_gbm	None	MR	T2 SAG SPACE: Processed_CaPTk
60173	upenn_gbm	None	MR	T2 SAG SPACE: Processed_CaPTk
60174	upenn_gbm	None	MR	T2 SAG SPACE: Processed_CaPTk
60175	upenn_gbm	None	MR	T2 SAG SPACE: Processed_CaPTk
60176	upenn_gbm	BRAIN	MR	T2 SAG SPACE: Processed_CaPTk

60177 rows × 4 columns

Figure 3. Illustration of the autocorrection pipeline for an example user input. The example demonstrates how the autocorrection pipeline recursively autoengineers the prompt to guide the LLM towards using the keyword SeriesDescription to filter different MRI sequences.

Keywords
Large language models; Natural language; Cohort discovery; Cancer Imaging; Human-computer interaction

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Pruning and Principal Component Analysis (PCA) on UNet++ for Segmentation of Kidneys and Cysts in Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Chetana Krishnan, Graduate Student, University of Alabama at Birmingham; Emma Schmidt; Ezinwanne Onuoha, MS; Michal Mrug, MD; Carlos E. Cardenas, PhD; Harrison Kim, PhD

Introduction
Training a convolutional neural network from scratch is time-consuming and requires substantial computational resources. Compression helps re-use a pre-trained model (PTM) (e.g., MobileNet) with little fine-tuned training and less memory. The compressed model is either trained on the same dataset as PTM or a different one, such as transfer learning. Pruning is a widely used compression technique that eliminates redundant weights (RW) from the PTM to reduce its size without compromising performance. However, it causes sparse connectivity in the network leading to irregular memory access patterns and reduced performance. We propose UNet++ (PTM) with PCA compression (PCAUNet++), which removes the unimportant filters from PTM instead of the weights preserving features accurately.

Hypothesis
PCAUNet++ outperforms UNet++ with pruning compression (prUNet++) in terms of preserving model performance while achieving higher compression ratios.

Methods
In prUNet++, the threshold is determined empirically within the range of 0.1 to 0.5 for each layer based on the weights and is iterated until RW from all layers is removed as a two-step process. PCAUNet++ identifies the number of principal components needed to explain 99.9% of the cumulative explained variance, allowing for the optimization of layer width. The depth is optimized based on when these significant dimensions start to contract, removing the need for iterations. The models were trained on T2-weighted MRI images of 95 ADPKD patients, utilizing 756 3D kidney images (604 for training, 76 for validation, and 76 for testing). Preprocessing, cropping, and slicing techniques were applied to generate 2D training samples, resulting in approximately 69,000 samples. The models were trained for 50 epochs using a patch-wise approach. Data augmentation techniques were employed to increase the training samples. Performance was evaluated using the Dice similarity coefficient (DSC), Hausdorff distance (HD), and Intersection over Union score (IoU).

Results
Figure 1 shows the representative kidney and cyst predictions. Figure 2 illustrates the network parameters of UNet++, prUNet++, and PCAUNet++. PCAUNet++ had about twice as fewer parameters as prUNet++ and about four times fewer parameters than UNet++. The training and inference time for PCAUNet++ was shorter than those for UNet++ and prUNet++ ($p < 0.0001$) (Table 1). The mean DSC remains the same for all three models. PCAUNet++ outperforms prUNet++ in cyst predictions.

Conclusion
PCAUNet++ performed lossless and uniform compression to preserve the maximum variance in the model.

Statement of Impact
PCAUNet++ can help clinicians reuse a model in clinical applications that involve training large datasets like ADPKD with less computational costs and time, improving diagnosis.

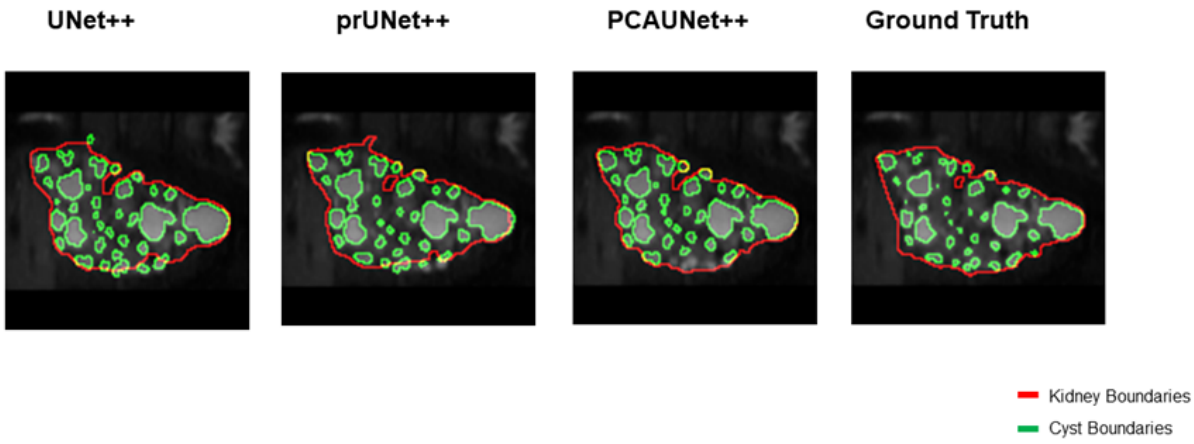


Figure 1. Kidney and cyst segmentation using UNet++, prUNet++, and PCAUNet++. Representative images showing segmentation on the test set with kidney (red line) and cyst (green line) boundaries determined by our semi-automatic method (ground truth, fourth column) and all three models (first to third columns). PCAUNet++ shows better kidney boundary localization but less cyst boundary localization. This is due to boundary distortion during compression. Post-processing techniques on predicted images like contrast enhancement or Gaussian filtering can improve accuracy.

Figure 1. Kidney and cyst segmentation using UNet++, prUNet++, and PCAUNet++. Representative images showing segmentation on the test set, respectively, with kidney (red line) and cyst (green line) boundaries determined by our semi-automatic method (ground truth, fourth column) and all three models (first to third columns). PCAUNet++ shows better kidney boundary localization but less cyst boundary localization. This is due to boundary distortion during compression. Post-processing techniques on predicted images like contrast enhancement or Gaussian filtering can improve accuracy.

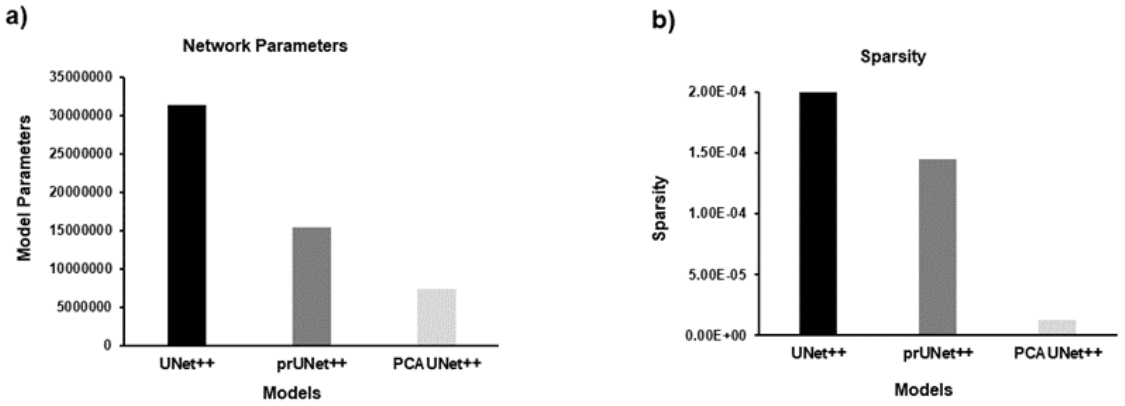


Figure 2. Network Architecture. a) Comparison between number of the model parameters of all the models. b) Comparison between the sparsity of all the models. PCAUNet++ shows fewer model parameters and lower sparsity meaning very fewer weights are close to zero and hence most of the weights contribute to the segmentation. Redundant weights from UNet++ causing high sparsity are removed to reduce sparsity by compression techniques.

Figure 2. Network Architecture. a) Comparison between the number of model parameters of all the models. b) Comparison between the sparsity of all the models. PCAUNet++ shows fewer model parameters and lower sparsity meaning very fewer weights are close to zero and hence most of the weights contribute to the segmentation. Redundant weights from UNet++ causing high sparsity are removed to reduce sparsity by compression techniques.

Model	ROI	IoU Score	HD (mm)	Test Dice	Min Test Dice	Max Test Dice	Training Time / Epoch (s)	Inference Time (min ± sec)
UNet++	Kidney	0.88±0.47	1.35±0.95	0.93±0.35	0.70	0.97	217±5	07±23
	Cyst	0.77±0.43	1.52±0.78	0.86±0.42	0.71	0.92	220±7	09±14
prUNet++	Kidney	0.87±0.45	1.42±0.98	0.93±0.34	0.71	0.97	110±2	05±12
	Cyst	0.73±0.49	1.65±0.89	0.84±0.43	0.52	0.91	124±4	06±34
PCAUNet++	Kidney	0.88±0.40	1.35±1.04	0.93±0.35	0.74	0.97	90±3	03±09
	Cyst	0.74±0.44	1.64±0.88	0.84±0.44	0.55	0.92	109±6	04±16

Table 1. Performance Metrics. Intersection over union (IoU) score, Haudsdorff distance (HD), and test Dice similarity score (DSC) of UNet++ and proposed models based kidney and cyst segmentation on the test set, together with its training time per epoch, and inference time to predict the boundaries of all test images. Fine-tuning a PTM takes less training time than training it from scratch.

*All three models are trained with the same dataset. Our future direction is to incorporate transfer learning.

Table 1. Performance Metrics. Intersection over union (IoU) score, Haudsdorff distance (HD), and test Dice similarity score (DSC) of UNet++ and proposed models based kidney and cyst segmentation on the test set, together with its training time per epoch, and inference time to predict the boundaries of all test images. Fine-tuning a PTM takes less training time than training it from scratch. *All three models are trained with the same dataset. Our future direction is to incorporate transfer learning.

Keywords
Pruning; Compression; Component analysis; UNet++; Segmentation; Polycystic kidney disease

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

SegViz: A federated Learning Framework to Train Multi-task Segmentation Models from Partially Annotated and Distributed Datasets

Adway U. Kanhere, MS, Software Engineer, University of Maryland School of Medicine; Pranav S. Kulkarni; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Introduction
Image segmentation is a foundational task in automated medical image analysis. However, datasets curated for training large-scale segmentation models are narrowly focused on only a subset of structures that may be simultaneously present in a patient's imaging study. This siloed approach would potentially result in hundreds of models that would need to be deployed in a clinical setup. To that end, we developed SegVIz, a collaborative learning platform where different groups working on different segmentation tasks can share knowledge with each other to train a global centralized model capable of performing all the tasks shared across the network

Hypothesis
Federated learning can aggregate knowledge from heterogeneous, distributed medical imaging datasets with partial annotations into a single 'global' model without the need to share data.

Methods
We developed SegViz to segment the Liver, Spleen, Pancreas, and Kidney using four CT datasets from the Medical Segmentation Decathalon challenge (N=131, 41, 282, and 210 scans, respectively), each containing annotations for only one of these organs (Figure 1). We used MONAI's 3D-U-Net implementation across all experiments. First, we trained four baseline 3D U-Net models, one on each dataset to represent the current state-of-the-art (a single model per task). Second, we trained a central aggregation model by combining all four datasets into a central repository. Finally, we trained the SegViz model using the FedBN algorithm as the aggregation strategy as it focuses on generalizing the global model to non-i.i.d data commonly encountered in medical imaging. We evaluated model performance using Dice Score on the internal validation and external Beyond the Cranial Vault (BTCV) dataset comprised of liver, spleen, pancreas, and kidney CT images (N=30).

Results
The SegViz model produced the best performance on the in-federation internal validation set as well as the out-of-federation BTCV test set, with dice scores of 0.93, 0.83, 0.55, and 0.75 for segmentation of liver, spleen, pancreas, and kidneys, respectively, significantly ($p < 0.05$) better (except spleen) than the dice scores of 0.88, 0.79, 0.46, and 0.64 for the baseline models (Figure 2, Table 1). The central aggregation model performed significantly($p < 0.05$) poorly on the test dataset with dice scores of 0.65, 0, 0.55, and 0.68.

Conclusion
Our results indicate that SegViz can be used to aggregate knowledge from distinct and distributed medical imaging clients without the need to share their data.

Statement of Impact
SegViz will enable different groups around the world in curating single-task datasets to collaborate and train large-scale clinically translational segmentation models.

	Mean Dice (SD)			
Models	**Liver**	**Spleen**	**Pancreas**	**Kidneys**
Baseline Liver	0.88 (0.15)	-	-	-
Baseline Spleen	-	0.79 (0.17)	-	-
Baseline Pancreas	-	-	0.46 (0.20)	-
Baseline Kidneys	-	-	-	0.64 (0.21)
Centrally agg	0.65 (0.14)	0	0.55 (0.18)	0.68 (0.21)
FedAvg	0.93 (0.02)	0.78 (0.14)	0.40 (0.20)	0.78 (0.12)
FedAvg + FT	0.93 (0.02)	0.73 (0.15)	0.44 (0.20)	0.76 (0.12)
FedBN	0.93 (0.01)	0.83 (0.14)	0.55 (0.17)	0.74 (0.11)
FedBN + FT	0.93 (0.01)	0.83 (0.15)	0.55(0.17)	0.75 (0.10)

Table 1. Mean Dice score performance of all the experiments on the out-of-federation BTCV dataset. The standard deviation values are in parentheses.

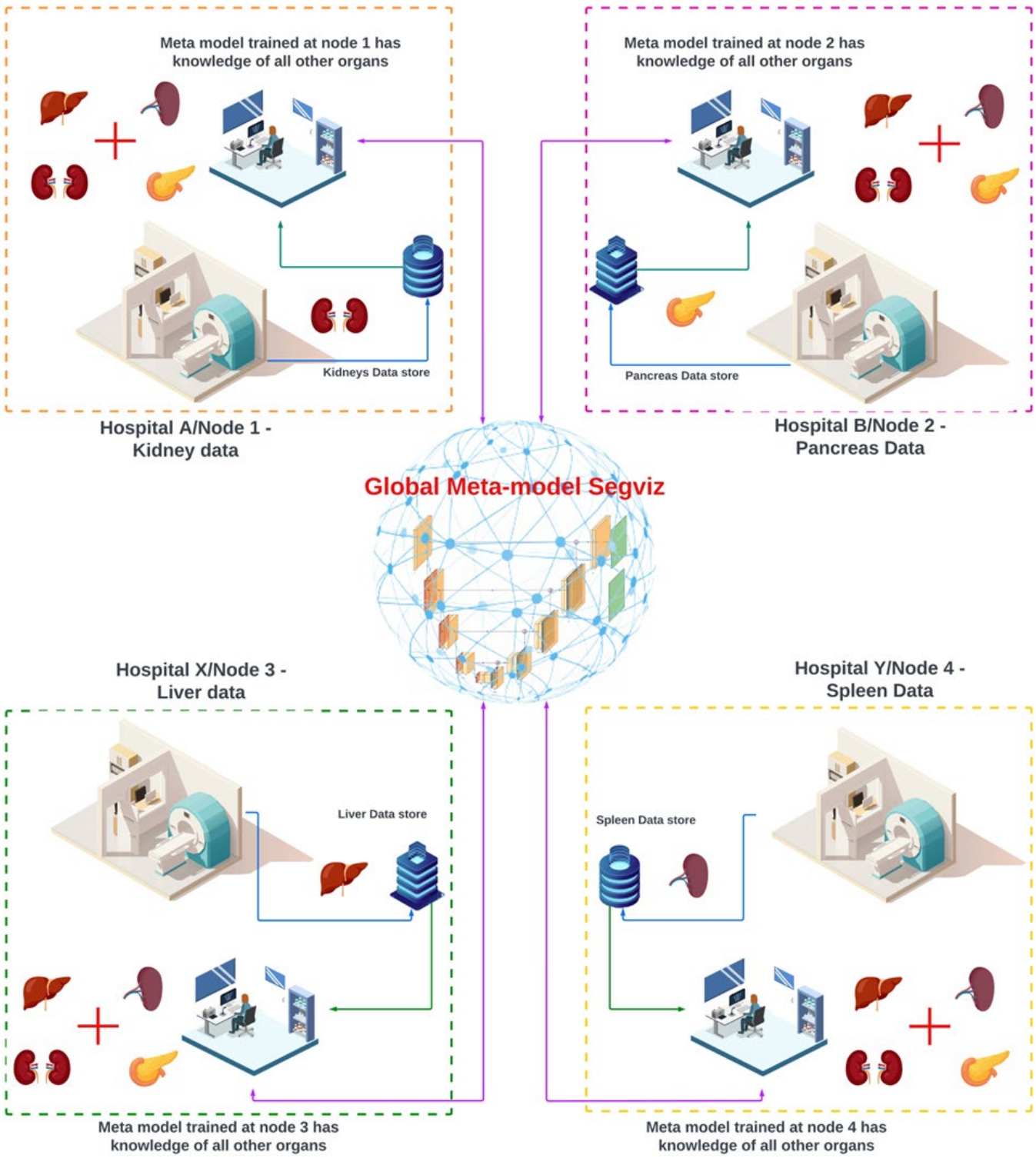


Fig. 1. Illustration of the proposed SegViz framework: Client nodes update the global meta-model where knowledge aggregation occurs after every 10 iterations of the local model. The weights of the global model are then shared with the client models allowing both nodes to share knowledge without sharing data.

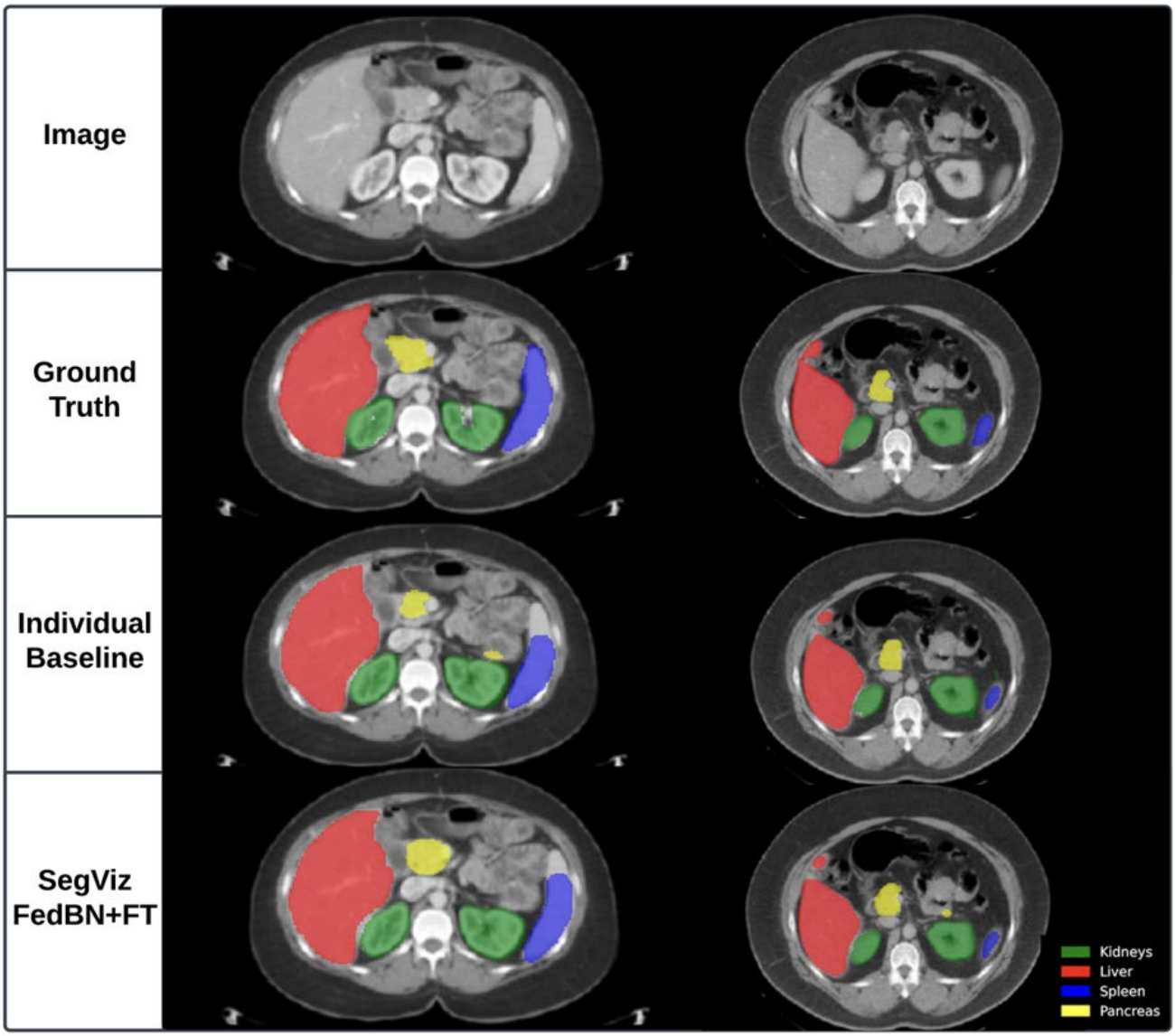


Fig. 2. A comparison of the ground truth segmentation masks with the masks generated by the baseline and SegViz models.

Keywords
Image segmentation; Federated learning; Partial annotations

Stimulated Raman Histology Image Reconstruction Using Weakly Supervised Generative Adversarial Networks

Sung Jik Cha, Medical Student, Western Michigan University; Yiwei Liu, MS; Esteban Urias, MD; Christian Freudiger, PhD; Todd Hollon, MD

Introduction

Stimulated Raman Histology (SRH) uses stimulated Raman scattering microscopy to create label-free histological images of biopsied tissues in a matter of seconds. SRH images have proven to be equivalent to standard H&E histology for intraoperative diagnosis, while drastically decreasing the time to image acquisition. One bottleneck in this process is the noisy degradation of the images due to microscope laser aberrations. This often requires repeated acquisition of the images, slowing down the operative workflow. Previous work has shown a weakly supervised U-Net is able to denoise noisy SRH images and outperforms denoising autoencoder and non-local means methods. We further build upon these efforts by introducing a method that uses generative adversarial networks with cycle consistency loss (CycleGAN) to further improve image quality without the need for paired data.

Hypothesis

Weakly supervised generative adversarial networks can reconstruct noisy, low quality brain tumor SRH images.

Methods

Training and validation datasets were generated from 4,589 high quality SRH patches of size 2*300*300 pixels from 368 patients. These patches were manually selected from a large database of 800,000 brain tumor SRH patches. A paired dataset was then created by applying additive and multiplicative Gaussian noise with randomly sampled variances to the high quality patches. A U-Net and a CycleGAN were trained on this paired data to transform the noisy, low quality patches into high quality patches. The CycleGan was modified to use a pixel-to-pixel L1 loss in addition to a cycle consistency loss to aid in image-to-image translation between the high and low quality patches.

Results

A test set of 2,134 paired high quality and low quality patches from 20 patients were obtained by repeatedly imaging the same tissue. Frechet Inception Distance (FID), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) scores were used to evaluate the reconstructed image quality. The CycleGAN outperformed the U-Net in FID and BRISQUE scores, while the U-Net outperformed the CycleGAN in SSIM and PSNR scores. Metrics are summarized in Table 1.

Conclusion

Both the U-Net and CycleGAN were able to reconstruct diagnostic cell features such as nuclear chromatin patterns from noisy, low quality SRH images. The CycleGAN outperformed the U-Net in unpaired metrics such as the BRISQUE and FID scores, and produced perceptually sharper reconstructions.

Statement of Impact

Fast, automated image reconstruction of noisy SRH images can improve intraoperative diagnostic accuracy and throughput.

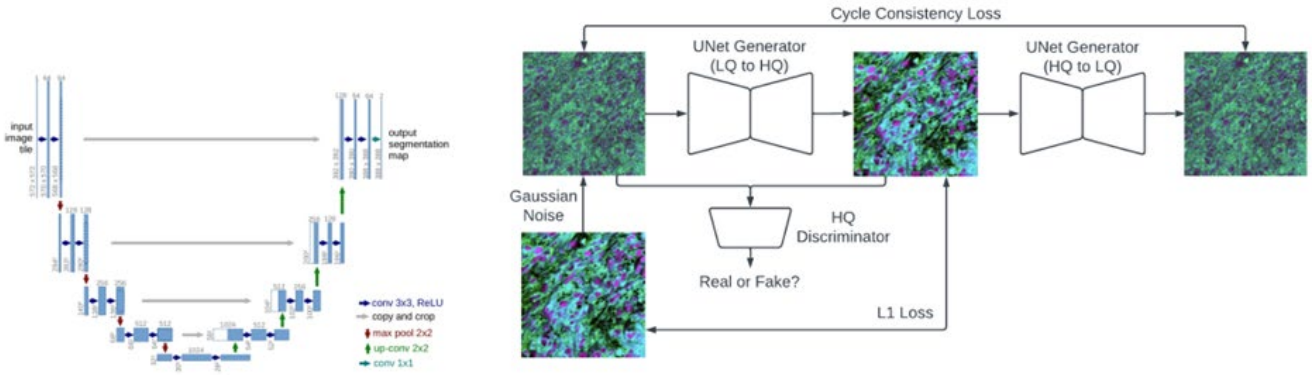


Figure 1. UNet (left) and CycleGAN (right) architectures. UNet consists of an encoder-decoder network with skip connections. CycleGAN consists of two generative adversarial networks, one for each direction of image-to-image translation. An additional L1 loss between the original high quality image and the reconstructed image was used in our model.

	SSIM	PSNR	FID	BRISQUE
HQ	1.000	80.000	0	24.470
LQ	0.228	12.654	47.558	54.776
UNet	**0.407**	**15.467**	39.816	29.886
CycleGAN	0.384	15.255	**19.131**	**24.090**

Table 1. Average SSIM, PSNR, FID, and BRISQUE scores. SSIM, PSNR, FID scores are computed between the high quality images and the low quality images, and between the high quality images and the images reconstructed by UNet and CycleGAN. Higher SSIM and PSNR scores, while lower FID and BRISQUE scores equate to better image quality.

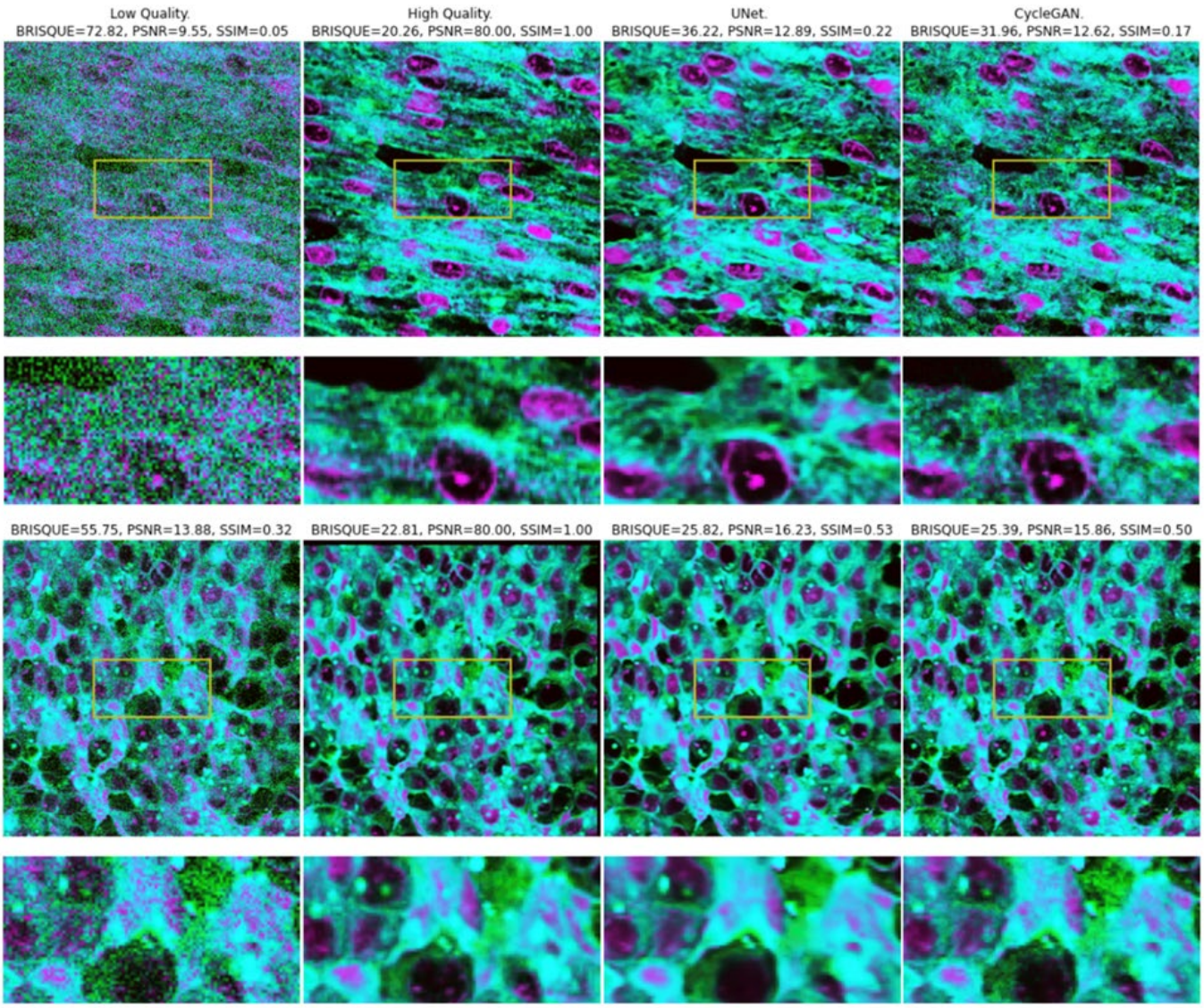


Figure 2. Perceptually, the U-Net tends to produce less noisy but smoothened reconstructions, while the CycleGAN produces slightly noisier but sharper reconstructions.

Keywords
Stimulated Raman Histology; Image Reconstruction; Denoising; Deep Neural Networks; Machine Learning; Generative Adversarial Networks

Unveiling Segmentation Errors: Enhancing Auto-Segmentation with ML Models Trained on Radiomic Features

Abishek Karki, PhD, Research Associate, University of Virginia; Victor G. Leandro Alves, PhD; Hashir N. Rashad, PhD; Jeffrey V. Siebers, PhD

Introduction

Target and organ-at-risk (OAR) delineation is crucial for effective radiation therapy planning. Errors in delineations can compromise treatment outcomes and increase the risks of complications. This study aims to mitigate the effects of erroneous delineations by developing a tool that automatically identifies errors in OAR segmentations, independent of a known ground truth.

Hypothesis

Gross delineation error in organ-at-risk delineations can be detected by inspecting the raw anomaly score from a Variational Autoencoder (VAE) of a delineation's radiomics.

Methods

The dataset for this study consisted of 210 Head and Neck cases obtained from publicly available sources. To ensure data quality, an in-house tool was used to detect and rectify common errors such as missing slices and ditzels. Over 109 radiomic features were computed from manually contoured CT images using the pyradiomics package. Additional features were added using in-house tools. Machine learning techniques, specifically an extra tree classifier and variational autoencoders from the PyOD package, were employed to detect anomalies within the dataset. To assess the accuracy of the methods, the dataset was enriched with auto-contours generated by both commercial and publicly available algorithms. In order to evaluate method robustness, deliberate perturbations were introduced by shifting the contours by 5 pixels in the x and y directions, creating pseudo-bad delineations. This allowed for the examination of technique performance under controlled variations in contour placement.

Results

The tool successfully detects simulated errors in organ-at-risk segmentations without reference of a known ground truth. For Parotid Left test data, an accuracy of 0.99 for |+ 5 mm shift| and 0.78 for |- 5mm shift| was achieved.

Conclusion

The results of our study supported our hypothesis, demonstrating that the implementation of the proposed methodology, specifically the use of a Variational Autoencoder (VAE) of radiomics features, effectively detected simulated errors in OAR delineations. By leveraging the VAE's capabilities, we achieved reliable and accurate identification of these errors, thus mitigating potential negative impacts on the quality of radiation therapy treatments.

Statement of Impact

The proposed delineation quality assurance tool could flag suspect manual or auto-delineations prior to their use in radiation treatment planning or other for other delineation uses, such as curating large datasets for auto-contouring training data.

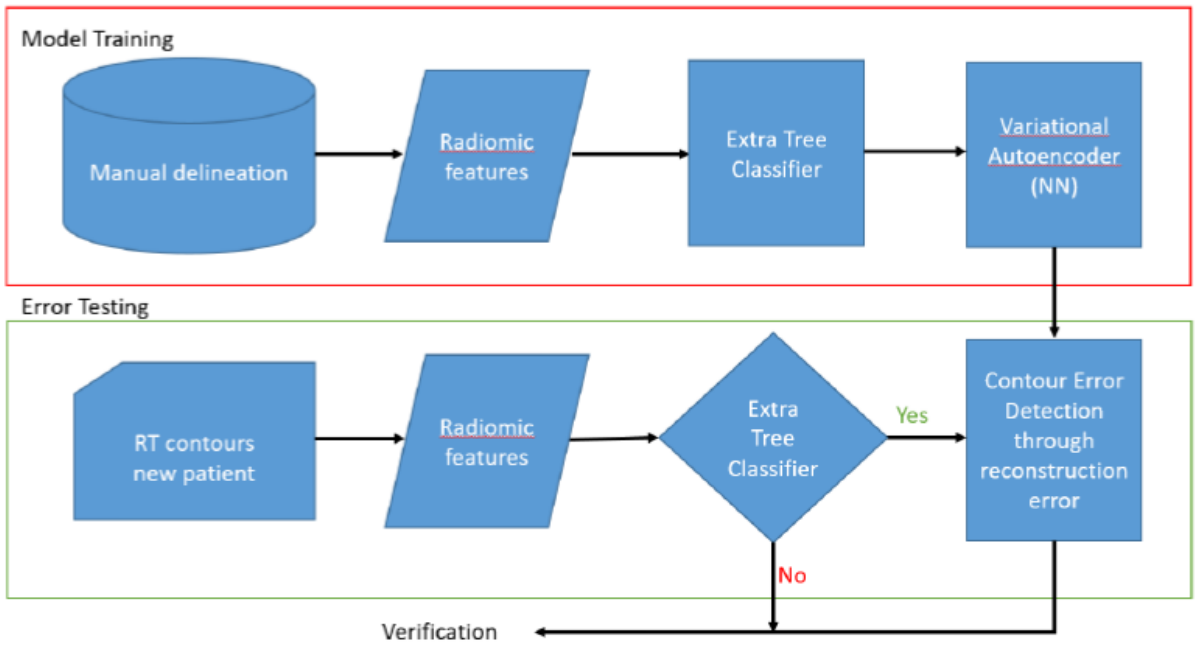


Figure 1 Visualization of the end-to-end methodology for detecting bad delineations in head and neck patients using machine learning and variational autoencoders.

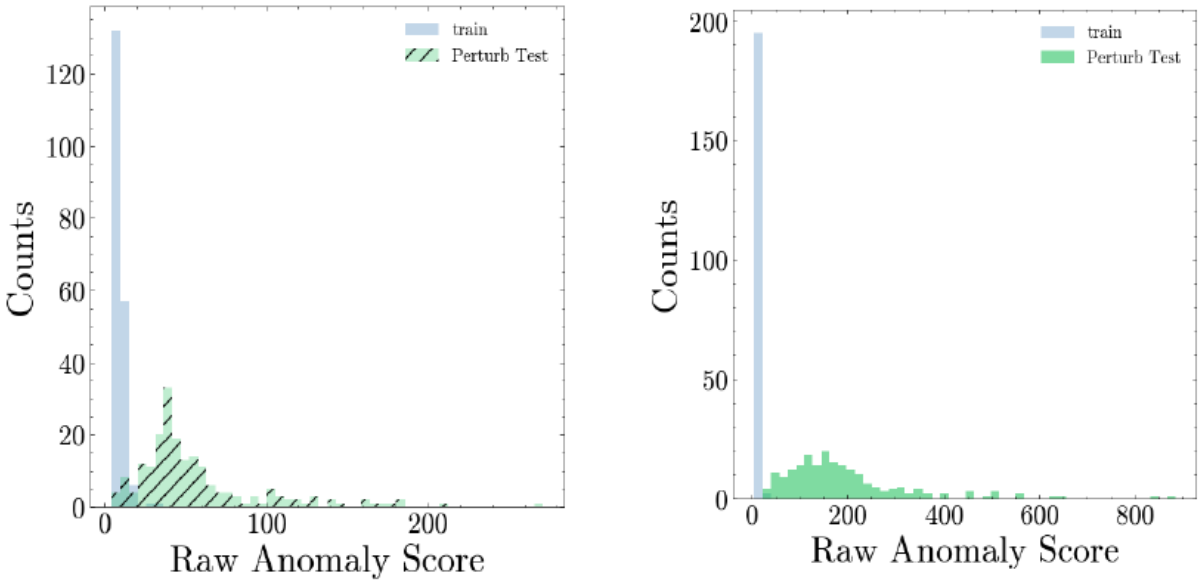


Figure 2 Comparison of raw anomaly scores for Parotid Left region of interest perturbations in the x and y directions. The left plot represents the anomaly scores when the ROI is perturbed by 5 mm in both the x and y directions, while the right plot displays the displays the scores when the ROI is perturbed by -5 mm. The light blue histogram in each plot represents the distribution of raw scores for the training samples

Keywords

organ-at-risk delineation; radiation therapy; error detection; machine learning; treatment planning; quality assurance

Using MONAI Pre-Trained Models for Colorectal Tissue Type Phenotyping: A Feasibility Study to Integrate Deep Learning Model Results using the Medical Extension OMOP CDM

Shijia Zhang, PhD Student, Johns Hopkins University; WooYeon Park, MS; Blake Dewery, PhD; Paul Nagy, PhD, CIIP, FSIIM

Background/Problem to be solved

The advent of deep learning models has transformed medical imaging analysis. However, effectively extracting coded imaging features from the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) imaging extension remains challenging for continuously phenotyping. OMOP-CDM is a standardized data model created by Observational Health Data Science and Informatics (OHDSI), designed to transform various medical images into a unified format. OHDSI, an open science collaborative community uses this model to perform reproducible, real-world analyses on observational health data with open-source software. The goal of our study is to integrate results derived from deep learning models into the OMOP data model. This allows us to combine deep phenotyping derived from imaging biomarkers to be integrated with health outcomes seen in the electronic medicine record.

Intervention(s)

We adopted a transfer learning paradigm, utilizing a pre-existing pathology tumor detection model from the Medical Open Network for AI (MONAI) and fine-tuned it using the PathMNIST Colon Pathology training dataset, stored in the OMOP-CDM image extension. The model, once fine-tuned, was applied to the testing PathMNIST dataset, generating tissue types which were subsequently recorded in the OMOP-CDM Conditions_occurrence table.

Outcome

The fine-tuned MONAI model outperformed the benchmark ResNet-18 model, with an improved overall AUC (0.995 vs. 0.989) and an increased accuracy (0.928 vs. 0.909). The model effectively phenotyped nine pathology tissue types and incorporated these findings into the OMOP-CDM without the need for parsing through unstructured reports. It's noteworthy that the model performed especially well on tissue types with higher prevalence.

Conclusion

Our study demonstrates that fine-tuned pre-trained machine learning models can successfully populate the OMOP-CDM with phenotypic evidence from pathology images, providing a more structured way to incorporate medical imaging findings into the OMOP-CDM without the process of sorting through unstructured reports. Moreover, this work shows the promise of federated learning across multiple medical institutions using the OMOP-CDM imaging extension, paving the way for widespread collaborative research and diagnostics.

Statement of Impact

This study stands to considerably advance the precision and accuracy of medical diagnoses and prognoses by refining phenotype definitions with cell-level or organ-level details. This enrichment can lead to more personalized medical solutions, aligning with the growing trend towards precision medicine. This study paves the way for efficient, machine learning-powered medical imaging analysis. Moreover, the potential for federated learning across multiple institutions could lead to a new era of collaborative diagnostics and research, enhancing patient outcomes.

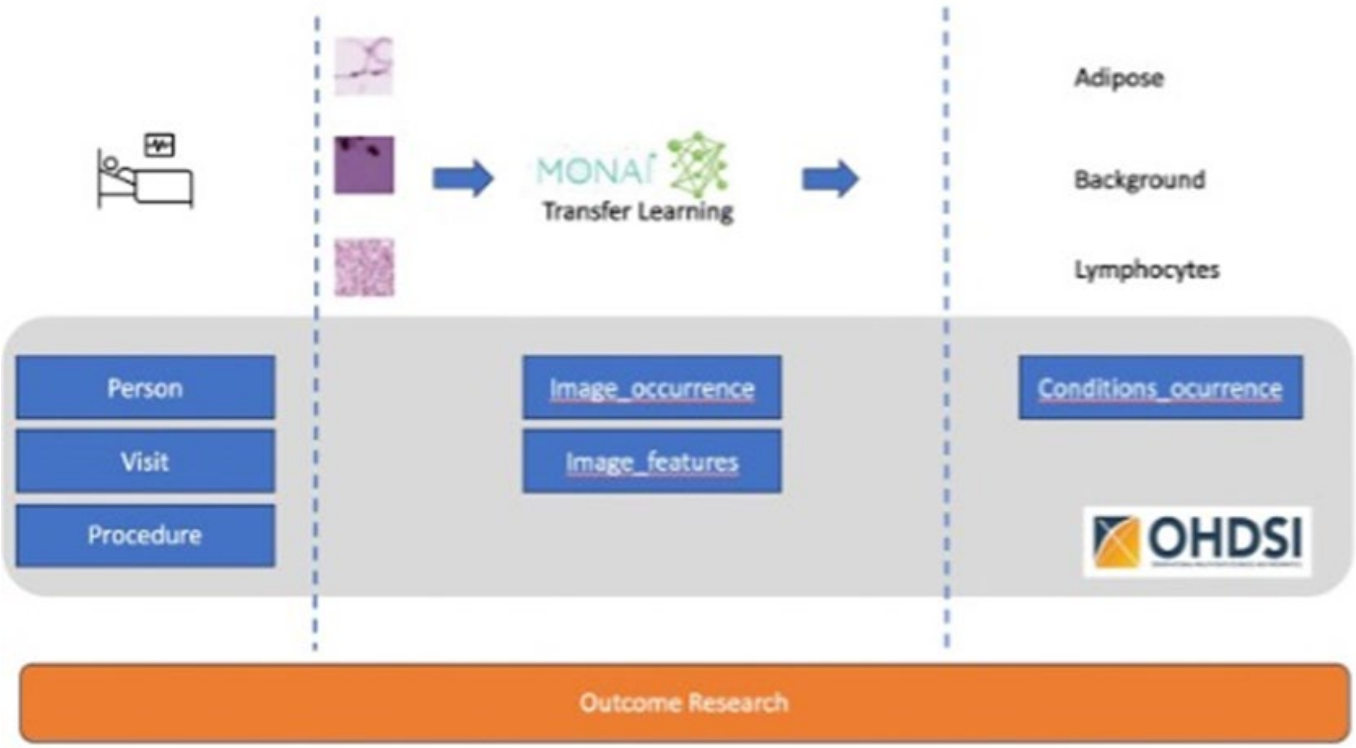


Figure1 - Visual Representations OMOP data encoding

	Precision	Recall	F1-Score	Num Cases
Adipose	0.96	0.95	0.95	1338
Background	0.95	1	0.98	847
Debris	0.79	0.9	0.84	339
Lymphocytes	0.97	1	0.98	634
Mucus	0.99	0.89	0.94	1035
Smooth Muscle	0.74	0.92	0.82	592
Normal Colon Mucosa	0.96	0.96	0.96	741
Cancer-Associated Stroma	0.95	0.53	0.68	421
Colorectal Adenocarcinoma Epithelium	0.94	0.97	0.96	1233

Table1 - Classification Performance by Tissue Types by MONAI's model

Keywords
MONAI; OMOP CDM; phenotyping; pathology tumor detection model; transfer learning; federated learning.

Scientific Abstract Presentations
Generative AI & General Applications in NLP

Date:
MON, OCT 2

Time:
8:45 AM – 10:15 AM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

Automatic Personalized Impression Generation for PET Reports Using Large Language Models
+ Xin Tie, MS, Graduate Research Assistant, University of Wisconsin-Madison
+ Muheon Shin, MD; Ali Pirasteh, MD; Ibrahim Nevein, MD; Zachary M. Huemann, MS; Junjie Hu, PhD; Steve Y. Cho, MD; Tyler J. Bradshaw, PhD

Hybrid Model for Whole-Body Synthetic CT Generation from TOF NAC PET Scans: Improved Accuracy and Attenuation Correction
+ Alan McMillan, PhD, Professor of Clinical Health Sciences, UW Health
+ Iman Z. Estakhraji, PhD; Tyler Bradshaw, PhD; PhD; Ali Pirasteh, MD

Improving the Readability of Patient-facing Information About Lung Cancer Using Large Language Models: ChatGPT, GPT-4 and Bard
+ Hana Haver, MD, Resident Physician, University of Maryland Medical Intelligent Imaging (UM2ii) Center & Massachusetts General Hospital
+ Jean Jeudy, MD; Cheng T. Lin, MD; Arlene Sirajuddin, MD; Paul H. Yi, MD, MS

Realistic Generation and Removal of Brain Tumoral Lesions on Multiparametric Brain MRI with Diffusion Models
+ Pouria Rouzrokh, MD, MPH, MHPE, Research Associate, Mayo Clinic AI Laboratory
+ Bardia Khosravi, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Sanaz Vahdati, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Synthesizing fMRI from MRI and EEG: A Deep Generative Approach
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Cooper Gamble

Utilizing Generative AI to Recognize Racial Disparities in Imaging Registries: A Step Toward Model Explainability
+ Bardia Khosravi, MD, MPH, MHPE, Postdoctoral Research Fellow, Mayo Clinic AI Laboratory
+ Pouria Rouzrokh, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Hillary W. Garner, MD; Doris E. Wenger, MD; Michael J. Taunton, MD; Cody C. Wyles, MD

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Automatic Personalized Impression Generation for PET Reports Using Large Language Models

Xin Tie, MS, Graduate Research Assistant, University of Wisconsin-Madison; Muheon Shin, MD; Ali Pirasteh, MD; Ibrahim Nevein, MD; Zachary M. Huemann, MS; Junjie Hu, PhD; Steve Y. Cho, MD; Tyler J. Bradshaw, PhD

Introduction
The process of deriving impressions from lengthy and complex findings in whole-body PET reports can be time-consuming. Large language models (LLMs) might accelerate workflows for interpreting physicians by automatically drafting impressions based on the findings. While there has been extensive research on summarizing radiology findings in X-ray, CT, and MRI reports, impression generation for whole-body PET imaging has received comparatively little attention.

Hypothesis
We hypothesized that LLMs fine-tuned on a large corpus of PET reports would accurately summarize PET findings and produce impressions that were deemed suitable for use within a clinical workflow.

Methods
Twelve LLMs, including 8 encoder-decoder models and 4 decoder-only models, were fine-tuned for PET report summarization using the standard teacher-forcing approach with the report findings as input and the original clinical impressions as ground truth. An extra token was used to encode the dictating physician's identity to allow the model to learn physician-specific reporting styles. Our corpus comprised 37,590 retrospective whole-body PET/CT and PET/MR reports collected at our institution between 2010 and 2022, with 4000 of the cases withheld for testing. To identify the best LLM for the task, 17 automatic evaluation metrics were benchmarked against the preferences of a nuclear medicine physician who scored and ranked the quality of impressions generated by 4 different language models on 200 cases. The evaluation metric that best correlated with physician preference was identified and used to compare the performances of the 12 fine-tuned LLMs and select the model that performed best overall. Three physicians then independently reviewed a total of 72 reports together with model-generated impressions, with each physician assessing 12 impressions originally dictated by themselves and 12 impressions originally dictated by other physicians. They evaluated the impressions based on 6 quality dimensions (3-point scale) and an overall utility score (5-point scale).

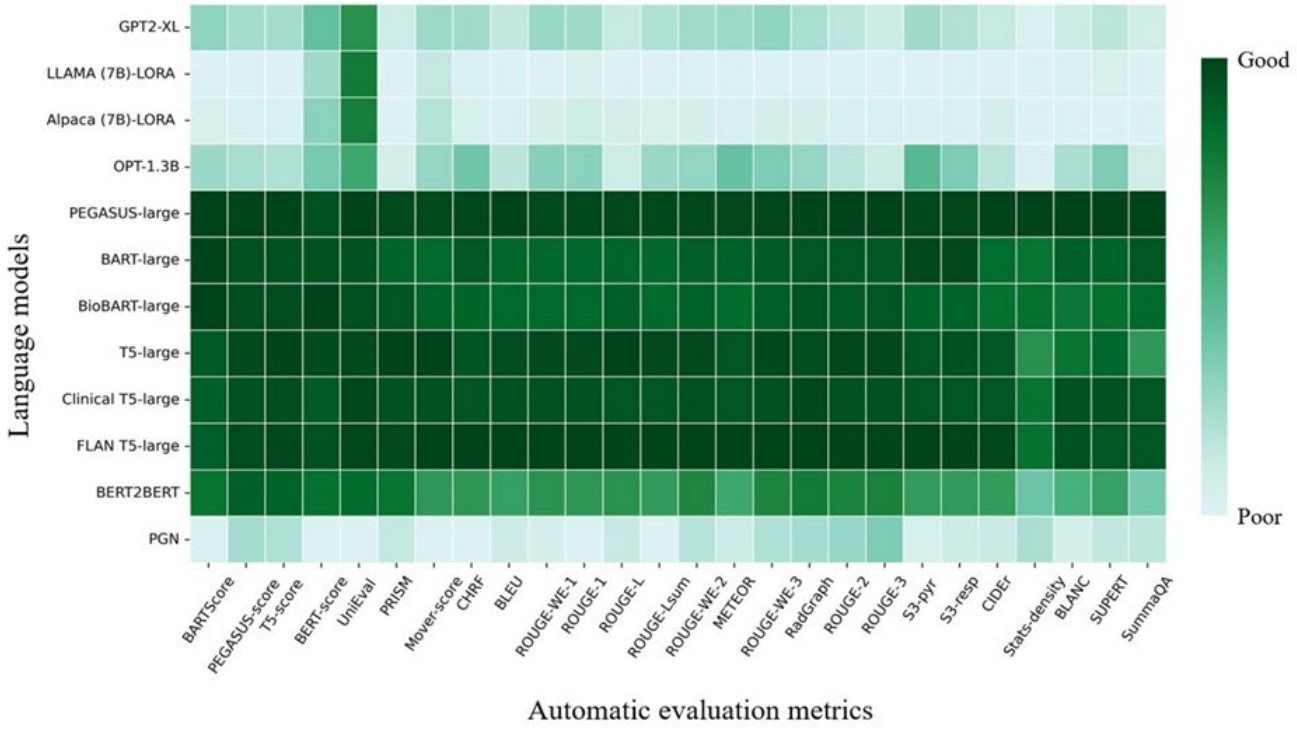
Results
Among all automatic evaluation metrics, BARTScore demonstrated the highest Spearman's correlation with physician preference (ρ=0.57). Based on BARTScore, the PEGASUS model was selected for expert evaluation. When physicians evaluated their own reports, 72% of the PEGASUS-generated impressions were clinically acceptable with either no changes or minor changes, with a mean overall utility score of 4.1 out of 5. On average, physicians preferred LLM impressions generated in their own style over impressions dictated by other physicians.

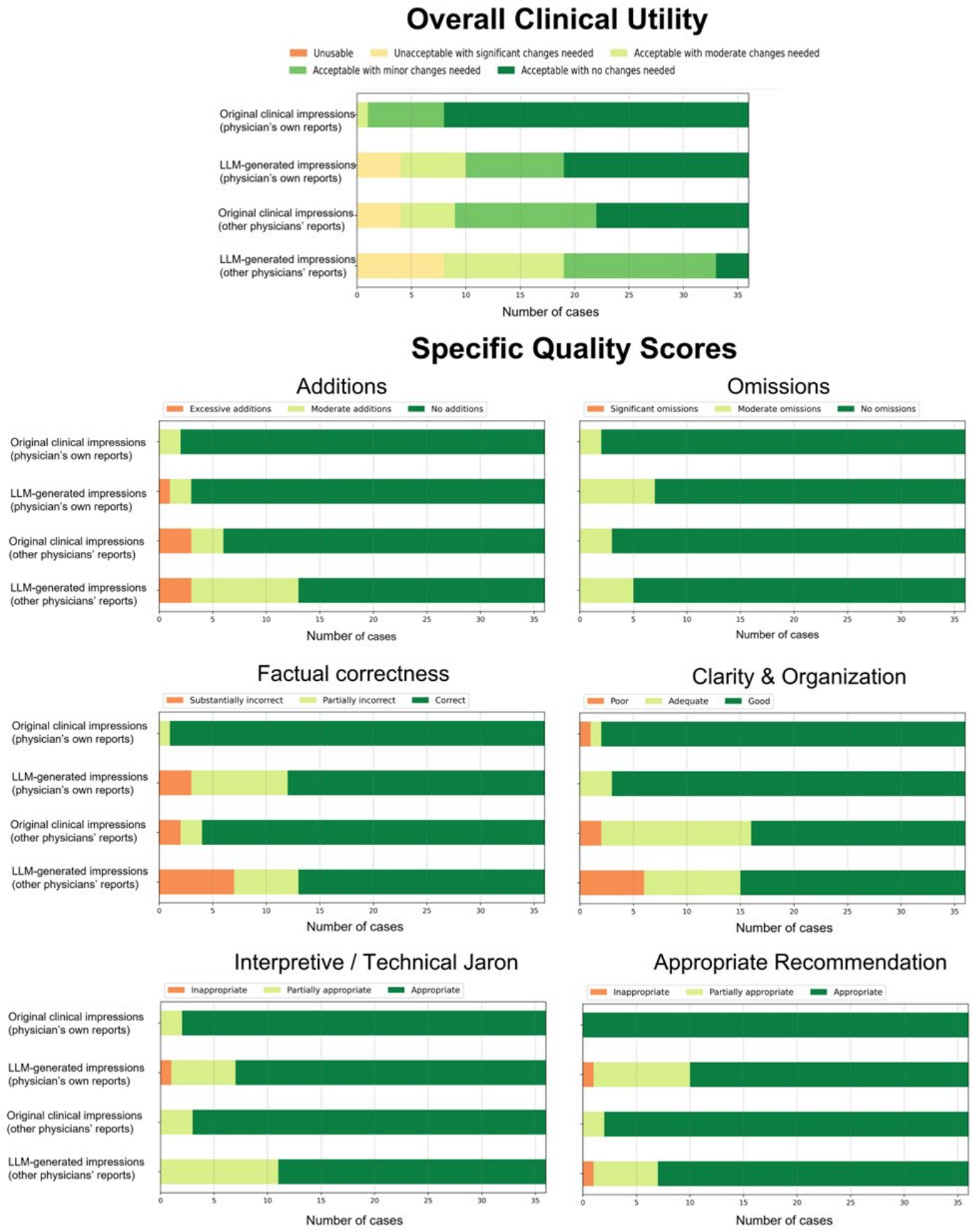
Conclusion
The top-performing model, PEGASUS, produced clinically-useful impressions in the large majority of cases.

Statement of Impact
LLMs have the potential to expedite and standardize PET reporting by automatically drafting impressions based on the findings.

The heatmap illustrates the performance evaluation of 12 language models fine-tuned for PET report summarization using 17 different automatic metrics. The X-axis displays the evaluation metrics arranged in descending order of correlation with physicians' preference, with higher correlations on the left and lower

correlations on the right. The values of each metric were normalized to allow comparison within a single plot. PEGASUS, BART, and T5 models performed similarly, with the PEGASUS model narrowly outperforming the other models when considering the metrics with the highest correlations.





Expert evaluation consisted of an overall clinical utility score and 6 specific quality scores. For the physician's own reports, 72% of the LLM-generated impressions were considered as acceptable with either no changes or minor changes needed.

Indication: [AGE] - year-old [SEX] with pulmonary nodule, presents for a staging FDG PET/CT examination.

Findings:
Background liver metabolic activity (SUV mean/ SUV max):
3.9/5.7 (PET/CT axial slice 155).
Background mediastinal blood pool metabolic activity (SUV mean/ SUV max):
3.1/3.9 (PET/CT axial slice 119).
Head/Neck:
No FDG avid cervical nodes are noted. Physiologic symmetric FDG uptake is present in the visualized portions of the brain, extraocular muscles, and salivary glands with no distinct focal abnormalities.
Chest:
Redemonstration of a subpleural oval-shaped solid nodule within the anteroinferior right upper lobe immediately superior to the right minor fissure, measuring approximately 1.2 x 1.3 cm in size, unchanged compared to [DATE]. This has mild associated FDG uptake (SUV max 1.8, axial slice 125). With reference to outside prior CT, there is suggestion of macroscopic fat within the nodule. No other pulmonary nodules are identified. No pleural effusion.
No FDG avid lymph nodes are noted in the axillae, hila, or mediastinum. Physiologic FDG uptake is present within the myocardium. No pericardial effusion. Mild calcified atherosclerotic plaque along the aortic arch and at the origins of the great vessels.
Redemonstration of a mildly lobulated nodular area of soft tissue within the subareolar right breast. This measures approximately 1.2 x 1.1 cm in size, unchanged compared to [DATE]. This does not appear to have associated FDG uptake.
Abdomen/Pelvis:
No FDG avid nodes or mesenteric lesions are noted. Heterogeneous FDG uptake is noted in the liver and spleen without focal abnormalities. The adrenal glands appear unremarkable. Mild uptake is noted along the bowel within the normal physiologic variation. Excreted radiotracer is present within the urinary collecting system and bladder. No hydronephrosis. Normal caliber abdominal aorta, with moderate calcified atherosclerotic plaque.
Musculoskeletal/Extremities:
No FDG avid skeletal lesions are noted. Infiltrated radiotracer is noted within the skeletal right upper arm antecubital region. There are no other areas of abnormal hypermetabolism within the visualized portions of the extremities.

Original clinical impression:

[1] FDG PET/CT demonstrates short-term stability of the solitary nodule within the right upper lobe, which has only minimal associated FDG uptake, not elevated above mediastinal blood pool. Given the lack of significant FDG avidity, and the suggestion of macroscopic fat within the nodule on prior diagnostic chest CT, a benign lesion such as a pulmonary hamartoma is a consideration. Could consider short-term follow-up diagnostic CT chest in 3-6 months to assess for interval change.

[2] No evidence of suspected FDG avid local nodal mediastinal/hilar or distant metastatic disease.

[3] Mildly lobulated nodular soft tissue in the subareolar right breast, without significant FDG avidity. Could correlate with prior mammography, if available. If not, screening mammography is recommended.

LLM-generated impression:

[1] FDG PET/CT demonstrates a 1.2 x 1.3 cm subpleural solid nodule within the anteroinferior right upper lobe with mild associated FDG uptake, which is nonspecific by PET criteria for an infectious/inflammatory process versus low-grade primary lung malignancy. Recommend correlation with prior and/or follow-up thoracic CT to assess for interval change in size.

[2] No evidence of FDG avid local nodal mediastinal/hilar or distant metastatic disease.

[3] Redemonstration of a mildly lobulated nodular area of soft tissue within the subareolar right breast, unchanged compared to [DATE]. This does not appear to have associated hypermetabolic activity.

This is an example case with the impression (bottom right) generated by PEGASUS based on the clinical PET/CT findings and indications. The bottom left section shows the original clinical impression for reference. The sentences with similar semantic meanings in the findings, original clinical impression, and LLM-generated impression are highlighted using identical colors. It is important to note that the recommendation (highlighted in magenta) is not stated in the findings and must be inferred by the LLM.

Keywords
Large language model (LLM); PET report summarization; Findings; Impressions

Hybrid Model for Whole-Body Synthetic CT Generation from TOF NAC PET Scans: Improved Accuracy and Attenuation Correction

Alan McMillan, PhD, Professor of Clinical Health Sciences, UW Health; Iman Z. Estakhraji, PhD; Tyler Bradshaw, PhD; Ali Pirasteh, MD

Introduction
We aimed to explore the performance of a hybrid model for synthesizing synthetic CT (sCT) images from TOF NAC PET scans. Accurate sCT synthesis plays a crucial role in various clinical applications, including radiation therapy planning and anatomical mapping.

Hypothesis
We hypothesized that the hybrid model, combining the SwinUNETR and UNET architectures, would enable accurate synthesis of sCT images from TOF NAC PET scans. We expected SwinUNETR to excel in capturing the bone region, while UNET would demonstrate superior performance in the soft-tissue region.

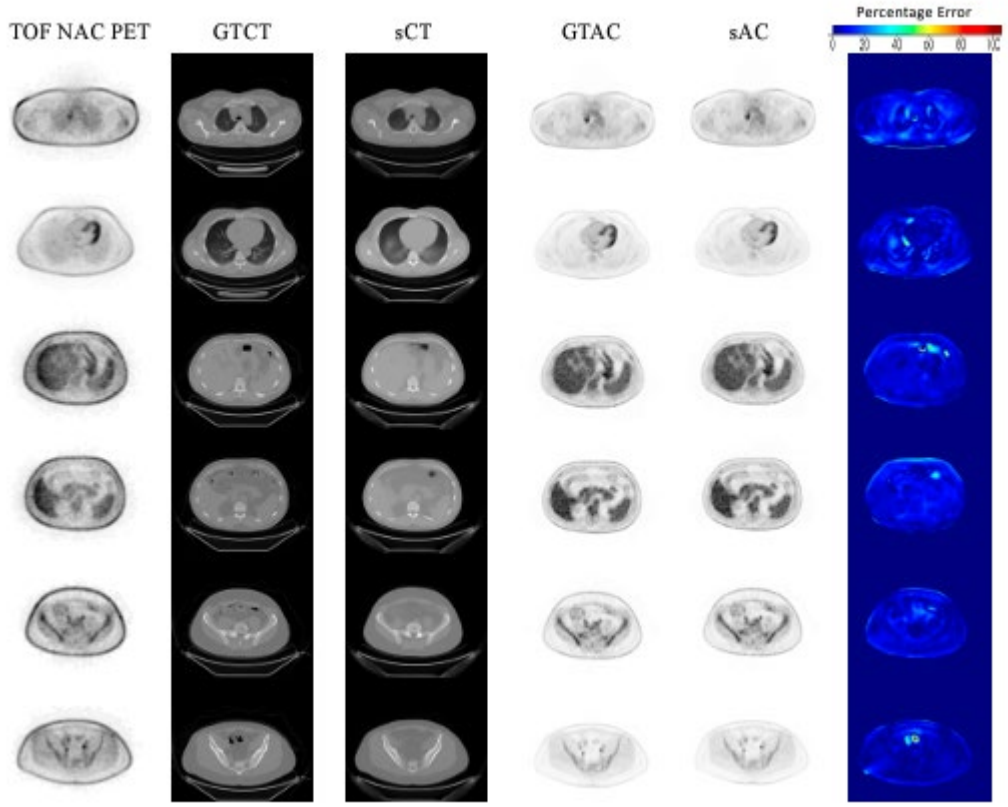
Methods
We employed the hybrid model that integrated the SwinUNETR and UNET architectures. SwinUNETR, known for its effectiveness in bone region capture, was utilized to synthesize the bone region of the sCT images. On the other hand, UNET, with its proficiency in handling soft-tissue characteristics, was employed to generate the soft-tissue region. To optimize the network parameters, we employed a multi-component loss function. Evaluation involved assessing the percentage error, Dice scores, and visual comparisons with ground truth CT and attenuation-corrected CT images.

Results
The results demonstrated the successful synthesis of sCT images resembling the ground truth CT scans. SwinUNETR exhibited superior performance in capturing the bone region, while UNET showcased excellent results for the soft-tissue region. The overall percentage error between the sCT and ground truth CT scans for the entire body was below 5%, indicating a high level of accuracy. Dice scores consistently exceeded 0.9 for both the whole body and brain regions, indicating strong agreement between the sCT and ground truth CT scans.

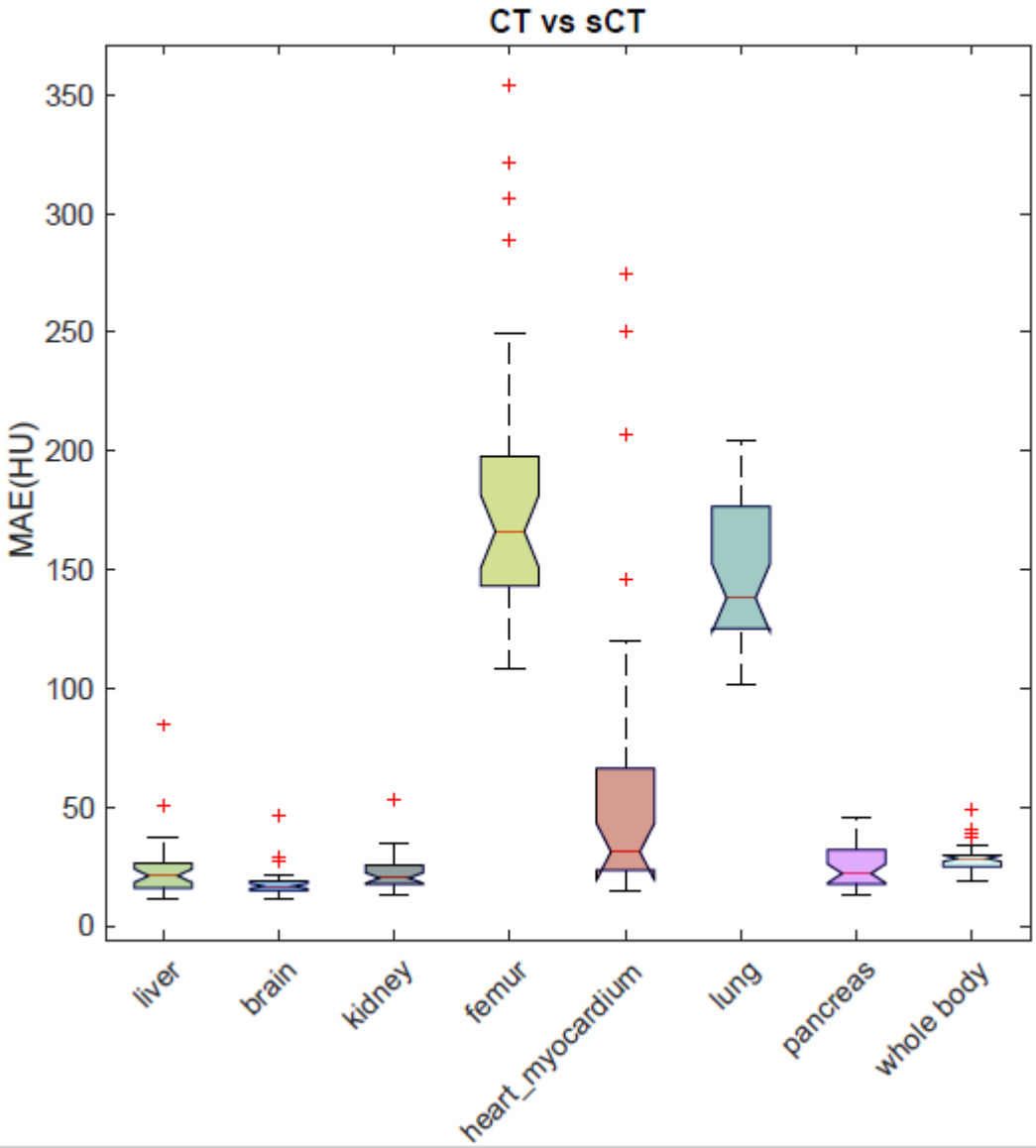
Conclusion
In conclusion, the hybrid model effectively synthesized accurate sCT images from TOF NAC PET scans. The combination of SwinUNETR and UNET architectures allowed for precise capture of the bone and soft-tissue regions, respectively. The low percentage error and high Dice scores demonstrated the model's accuracy in generating sCT images that closely resembled the ground truth CT scans.

Statement of Impact
This study demonstrates the potential of synthetic CT (sCT) generated from TOF NAC PET scans for accurate attenuation correction. By utilizing sCT alongside the ground truth CT (GTCT) images, named sAC and GTAC respectively, we observed a striking resemblance between sAC and GTAC. This similarity indicates that sCT can effectively serve as a substitute for GTCT in the attenuation correction process of TOF NAC PET imaging. The ability to use sCT for attenuation correction has significant implications in clinical practice, offering a more streamlined and efficient workflow by eliminating the need for additional CT scans.

Comparing TOFNAC PE, GTCT, sAC, GTAC, and sAC: Analyzing Slice Performance and Percentage Error

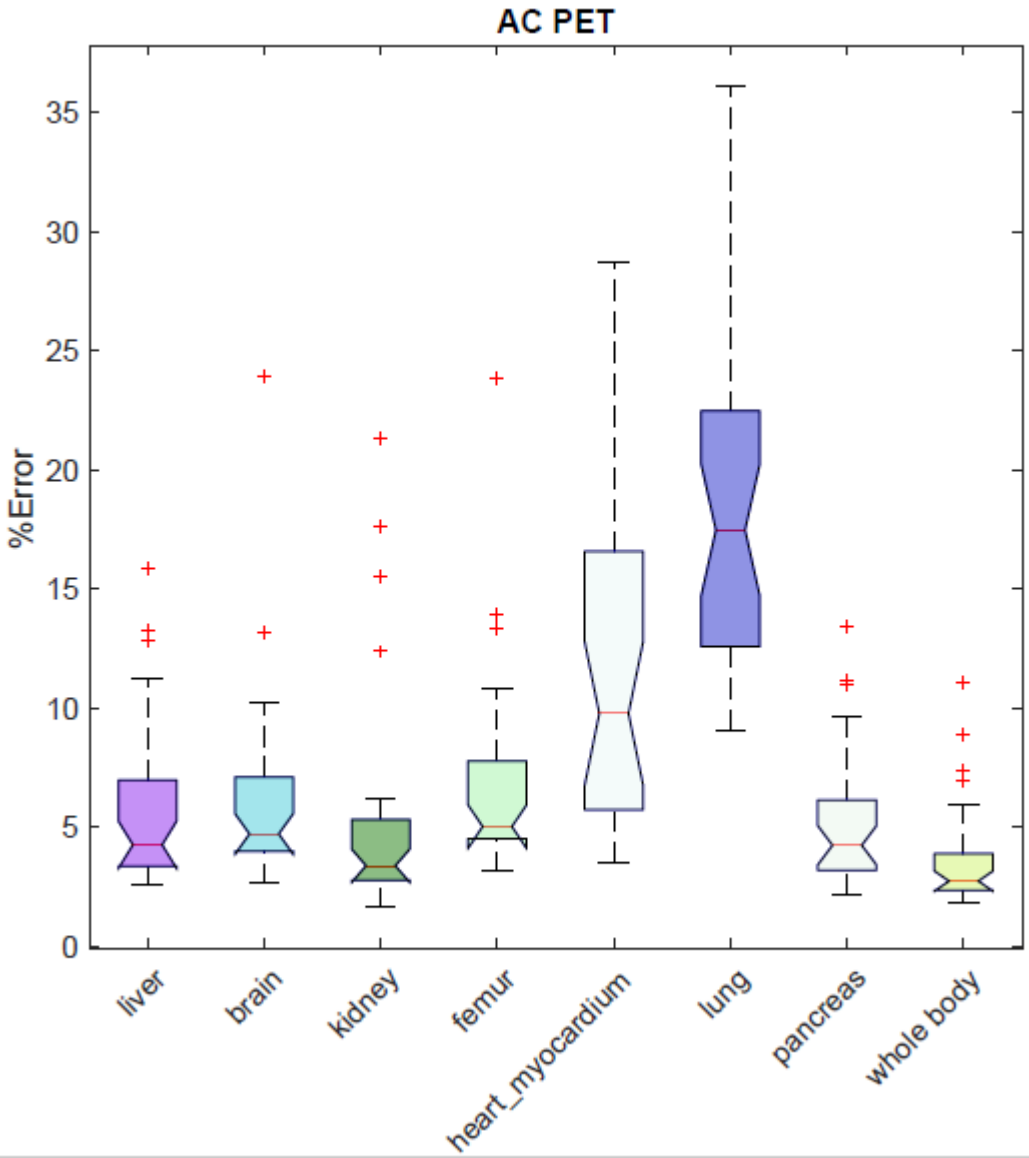


HAPYGCNX-1580863-2-ANY.pdf
Evaluation of Deep Learning Model Performance. Utilizing the TotalSegmentator package, we segmented various body parts from ground truth CT images of the test set. The resulting masks represented anatomical structures such as the liver, brain, kidney, femur, heart myocardium, lung, pancreas, and whole body. These masks served as binary maps indicating presence or absence of each body part within the corresponding CT image. Applying these masks, we assessed the performance of our deep learning model on each region of interest (sCT), comparing its predictions to the corresponding segmented masks. Mean Absolute Error (MAE) values between synthetic CT (sCT) and ground truth CT (GTCT) images were calculated and plotted in the above figure.



HAPYGCNX-1580863-3-ANY.pdf

Evaluation of Attenuation Correction Performance in PET Imaging for Different Organs. The figure illustrates the percentage error calculated for different organs when comparing the synthetic attenuation-corrected PET (sAC) images with the ground truth attenuation-corrected PET (GTAC) images. The percentage error calculated using Eq. 4, quantifies the deviation between sAC and GTAC images, providing insights into the accuracy of the synthetic attenuation correction method for different organs. This figure presents the assessment of attenuation correction performance in positron emission tomography (PET) imaging across various organs. Statistical analysis was conducted to evaluate the accuracy and consistency of the attenuation correction results. The findings provide insights into the effectiveness of attenuation correction techniques for precise organ-specific PET imaging.



Keywords
Synthetic CT; PET to CT synthesis; attenuation correction; TOF NAC PET; SwinUNETR; UNET

Improving the Readability of Patient-facing Information About Lung Cancer Using Large Language Models: ChatGPT, GPT-4 and Bard

Hana Haver, MD, Resident Physician, University of Maryland Medical Intelligent Imaging (UM2ii) Center & Massachusetts General Hospital; Jean Jeudy, MD; Cheng T. Lin, MD; Arlene Sirajuddin, MD; Paul H. Yi, MD, MS

Introduction
Access to information remains a modifiable contributor to disparities in lung cancer screening. With the increasing attention towards using large language models in generating health information has shown early promise, little has been reported about the capability of these models to improve readability of health information and its accessibility to the general public. Here, we assessed how 3 large language models (LLMs) perform in summarizing and simplifying health information about lung cancer prevention and screening.

Hypothesis
ChatGPT can provide accurate recommendations for lung cancer questions that are written at high readability levels. LLMs can improve readability of these recommendations to be more appropriateness for patients.

Methods
This retrospective study was designated as non-human subjects research by our IRB. We assessed the readability of answers to common questions about lung cancer prevention and screening by ChatGPT. We asked ChatGPT, GPT-4 and Bard to simplify the same set of answers, which were assessed for language complexity (Flesch Reading Ease) and readability on 5 established scales: Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, Automated Readability Index, and Simple Measure of Gobbledygook. Statistical analysis utilized paired t-testing between readability scores from the original and simplified answers for each model. Simplified answers were blindly rated for clinical appropriateness by 3 fellowship-trained cardiothoracic radiologists.

Results
The baseline answers to questions generated by ChatGPT in response to questions related to lung cancer had an overall mean language complexity of 49.7 and an overall mean readability of grade 12.6. Following simplification of these answers by ChatGPT (, GPT-4 and Bard, the overall mean language complexity and readability were found to be improved when each was compared to the score of the original text (Figure 1). Upon blinded assessment of clinical appropriateness of the simplified answers, our board-certified cardiothoracic radiologists determined that they were clinically appropriate 84% (ChatGPT), 79% (GPT-4) and 95% (Bard) of the time.

Conclusion
Lung cancer information provided by ChatGPT is complex and difficult to read. ChatGPT, GPT-4 and Bard demonstrate the capability to simplify the language of lung cancer information to a level more accessible for the general public, though the non-trivial number of inappropriate or inconsistent answers of the simplified content imply that further study is required.

Statement of Impact
ChatGPT generates responses to questions about lung cancer that are difficult to read. ChatGPT, GPT-4 and Bard can simplify this text with varying degrees of clinical appropriateness.

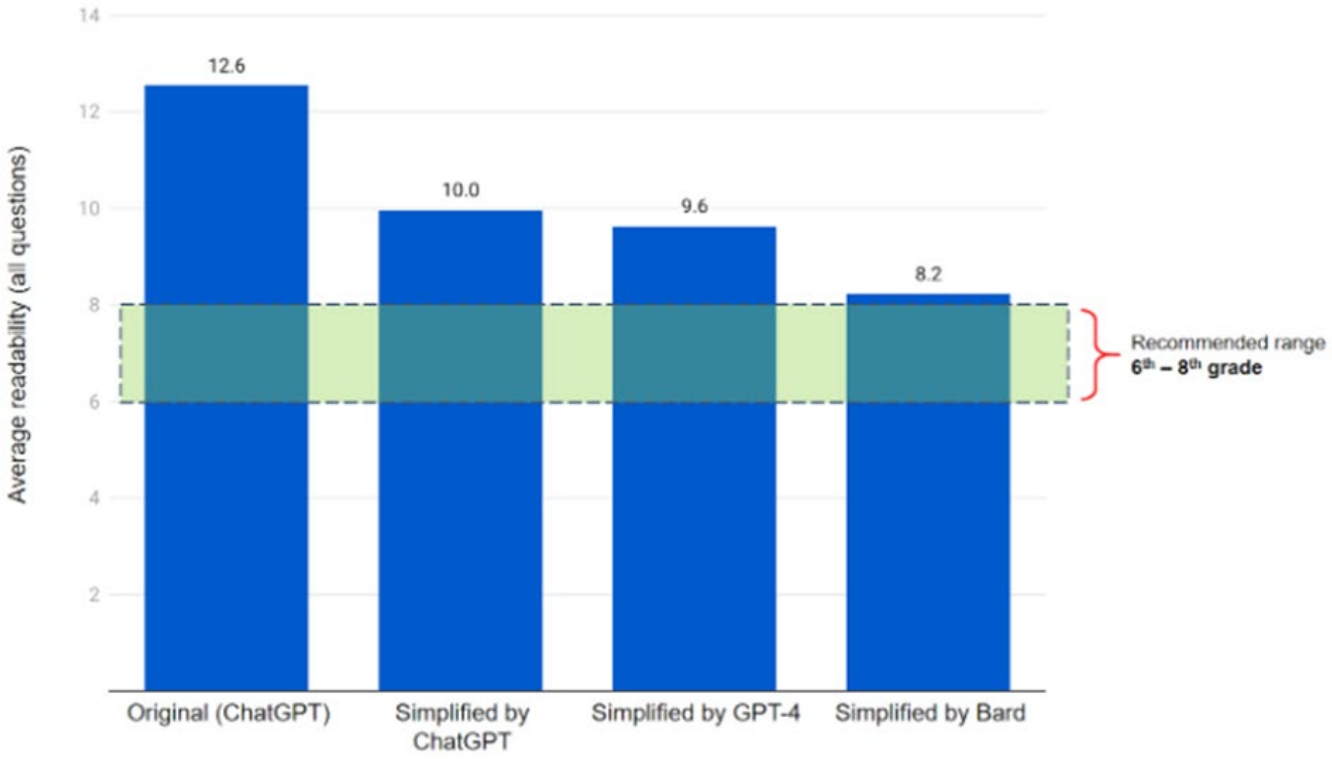


Figure 1. Average readability scores of the original responses (n=57) to 19 questions about lung cancer prevention and screening generated by ChatGPT and those simplified by different models (ChatGPT, GPT-4 and Bard).

Keywords
Large language models; Lung cancer; Patient education; Readability; Health literacy

Realistic Generation and Removal of Brain Tumoral Lesions on Multiparametric Brain MRI with Diffusion Models

Pouria Rouzrokh, MD, MPH, MHPE, Research Associate, Mayo Clinic AI Laboratory;
Bardia Khosravi, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Sanaz Vahdati, MD;
Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction
Despite the ever-increasing interest in applying deep learning (DL) models to medical imaging, the scarcity and imbalance of medical datasets can severely impact the performance of DL models. The generation of synthetic data that might be freely shared without compromising patient privacy is a well-known technique for addressing these difficulties. Our study introduces a DL algorithm that can add or remove brain glioma lesions from T1, T1-contrast enhanced (T1-CE), T2, or FLAIR MRI sequences.

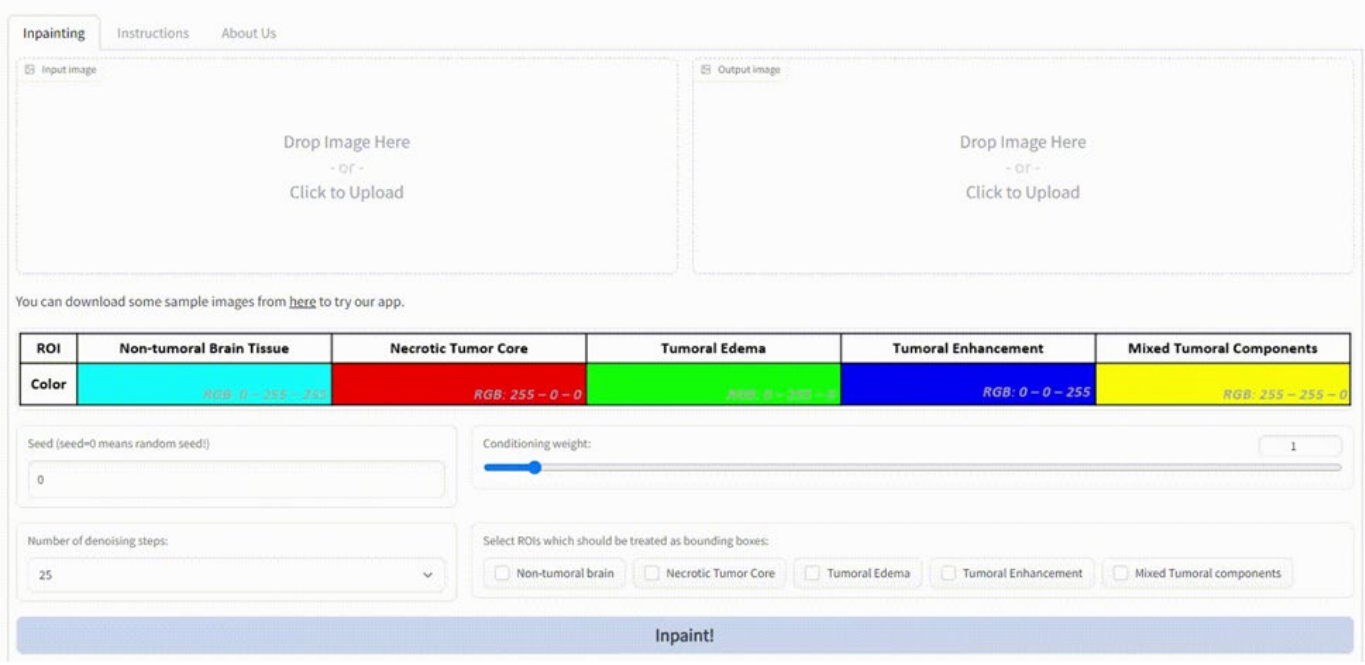
Hypothesis
We hypothesize that our tool can be employed to modify existing MRI images from patients with and without brain tumors, thereby generating synthetic but realistic data for subsequent DL training.

Methods
We developed an inpainting denoising diffusion probabilistic model (DDPM) that accepts a region of interest (ROI) and a generation mode from the user, then inpaints the ROI based on the chosen mode using classifier-free conditioning guidance. More specifically, the model can be instructed to generate (1) tumoral lesions with user-specified boundaries for necrotic tumor core, tumoral enhancement, and/or edema, (2) a tumoral lesion with no user input on its components, (3) a bounding box area of brain tissue that includes a tumoral lesion and surrounding normal tissue with compression effects, and (4) to replace tumoral lesions with normal-appearing brain tissue. The model was trained on 1251 skull-stripped brain MRI studies from the Brain Tumor Segmentation Challenge (BraTS) 2021.

Results
Due to the non-deterministic nature of DDPMs, conventional inpainting metrics (SSIM/PSNR) are unsuitable for their evaluation. Instead, a board-certified neuroradiologist with >30 years of experience reviewed 100 real and 100 synthetic images from our test set. The radiologist's accuracy in identifying synthetic imaging ranged from 43.1% to 51.6% across different tasks and views. No significant differences were found in the radiologist’s choices between real and synthetic data. A graphical user interface was also created to demonstrate the tool’s capabilities.

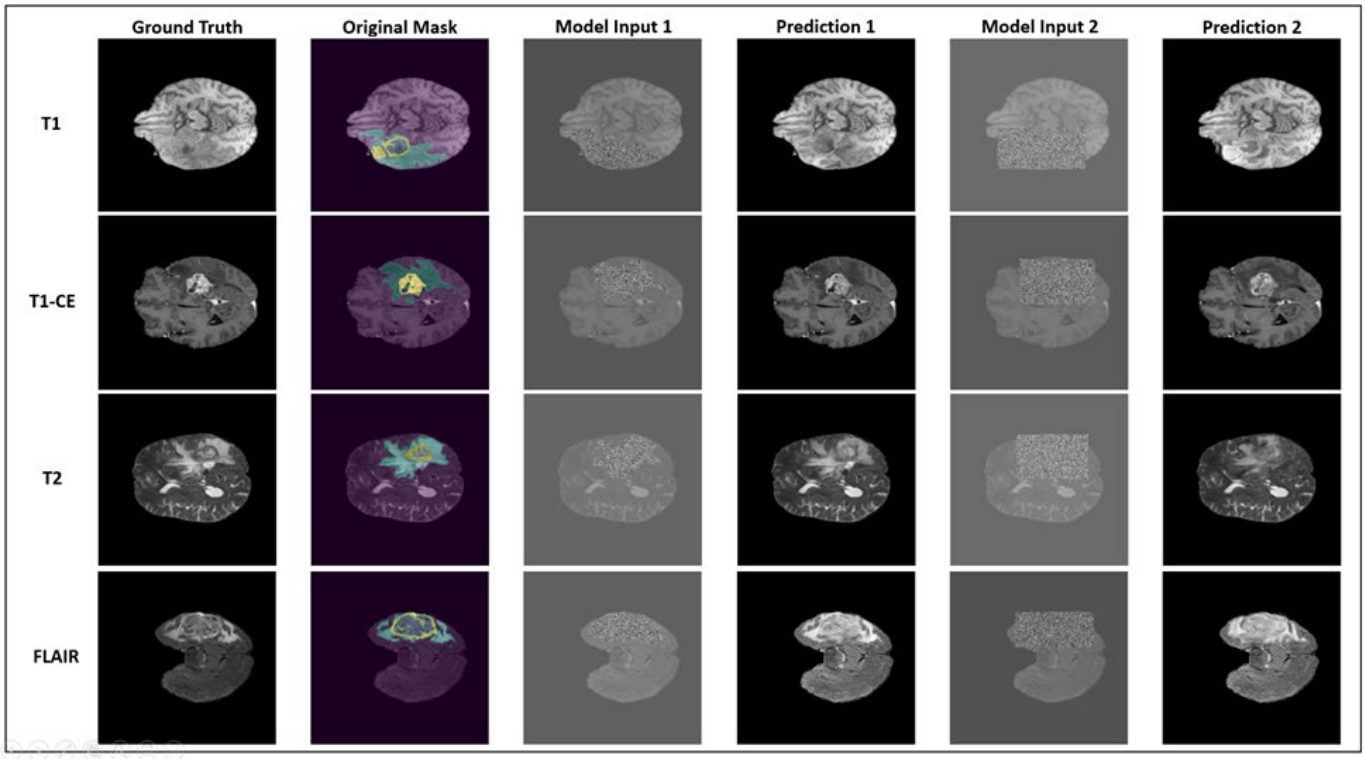
Conclusion
We developed a diffusion-based inpainting model that can add brain tumoral lesions with user-specified features to brain MRI imaging, or inversely, replace a tumoral region with normal brain anatomy. Our generated imaging looked as realistic as real brain MRI imaging in expert evaluation.

Statement of Impact
The scarcity and imbalance of medical datasets can significantly hinder the performance of DL models. Our model serves as a proof-of-concept that illustrates the potential of inpainting DL models to add synthetic brain tumoral lesions to existing normal brain MRI studies and therefore make an altered version of the existing data available for future DL training.

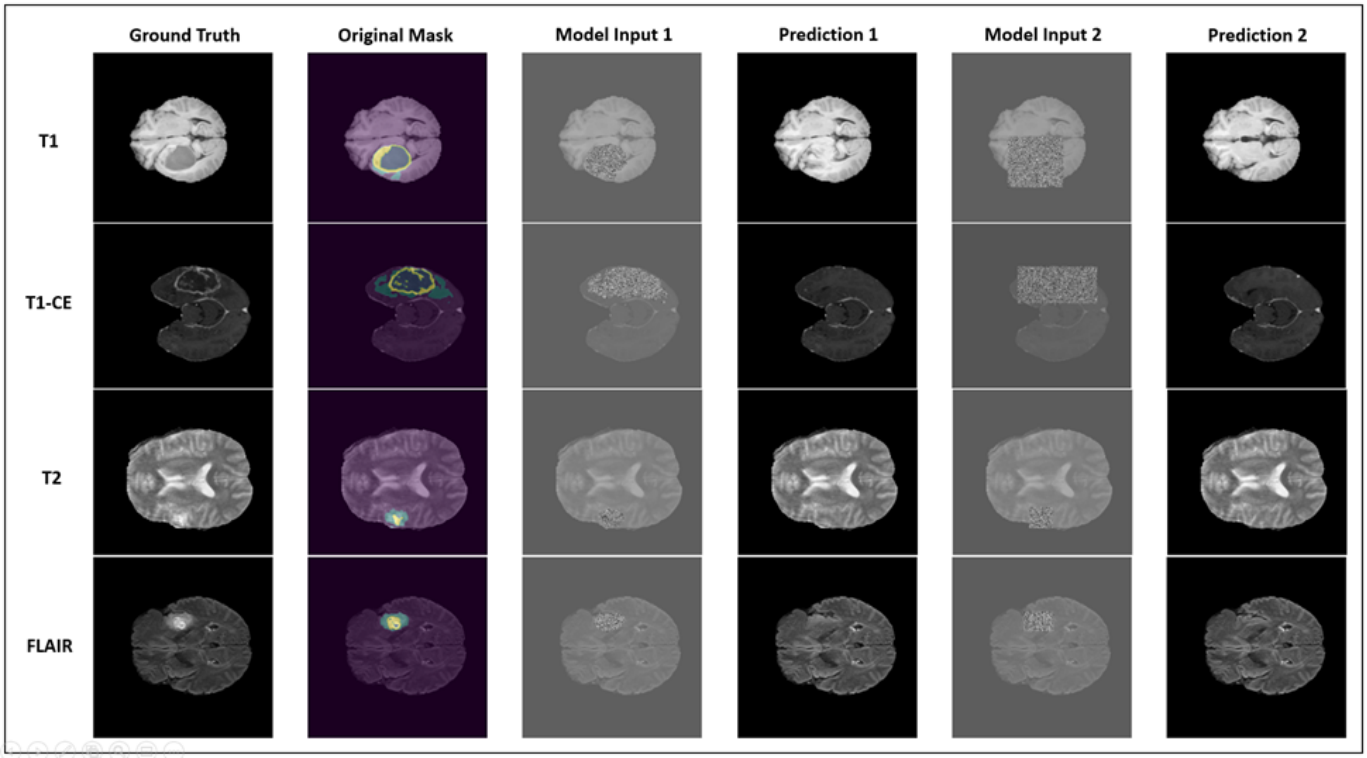


Demo link: https://www.abstractscorecard.com/uploads/Tasks/upload/20687/HAPYGCNX-1578969-1-ANY.gif

Demo of a graphic user interface (GUI) that demonstrates our inpainting model. The colors in the figure represent different tasks that the model can do at the same time; e.g., the blue color in the demo signals the model to replace a region with normal brain tissue, and the yellow color signals the model to replace the region with a tumor without specifying its components. Behind the scenes, each color is encoded as a conditioning variable and fed to the model in addition to the input imaging data.



Generating tumoral lesions with user-specified regions of interest (ROIs) for necrotic tumor core, tumoral edema, and tumoral enhancement. In each instance, prediction 1 was done with a free-form input ROI and prediction 2 was done with a bounding box input ROI.



Generating tumor-less (apparently normal) brain tissues instead the regions of interest (ROIs) for tumoral lesions. In each instance, prediction 1 was done with a free-form input ROI and prediction 2 was done with a bounding box input ROI.

Keywords
Brain tumors; Glioma; Inpainting; Generative AI; Diffusion models; DDPM

Synthesizing fMRI from MRI and EEG: A Deep Generative Approach

Cooper Gamble, Undergraduate Research Intern, Mayo Clinic; Shahriar Faghani, MD

Introduction
Functional Magnetic Resonance Imaging (fMRI) is a powerful technique for time series analysis of neurological and psychiatric pathologies, but acquisition cost and time is often an inhibitor for its widespread use. Magnetic Resonance Imaging (MRI) and Electroencephalogram (EEG) are common imaging techniques for structural and electrical analysis, respectively. In addition, their comparative acquisition cost and time is markedly more practical than that of fMRI. In this study, we demonstrate that MRI and EEG can be combined in a multimodal deep learning (DL) approach to generate high-fidelity, low-cost fMRI volumes.

Hypothesis
Data acquired through fMRI can be alternatively obtained by applying deep learning to EEG and MRI.

Methods
We used a publicly available dataset (Lioi, G., et al. 2019) containing structural MRIs with simultaneous 64-channel EEG and fMRI during right-hand motor imagery and neurofeedback. We skull-stripped and coregistered each scan with FMRIB's FSL tool and University College London's Statistical Parametric Mapping. We developed and trained a 3D U-Net on 7,901 samples, each of which aimed to predict one time step in the target fMRI scan. We prioritized the preservation of physiological data from MRI scans and injection of electrical data from EEGs by attempting to compress and reconstruct MRIs while feeding EEG data to our model at each encoding and decoding step. Our objective function was mean squared error (MSE), and we reported the structural similarity index measure (SSIM) of our model's predictions and the corresponding ground truths. Both of these metrics are recognized as indicators of generative performance in DL.

Results
After training for 82 epochs, our model achieved an MSE of 0.260 and an SSIM of 0.337.

Conclusion
We developed and trained a DL model to produce high-fidelity fMRI samples from EEG and MRI. We present an approach to reduce the temporal and fiscal cost of acquisition for fMRI by combining two practical modalities. Our next steps include pruning our network architecture to improve generative metrics and validating our methods on downstream tasks to demonstrate generalizability.

Statement of Impact
fMRI is a costly imaging technique with a demanding acquisition but of high clinical value. We present a DL-based approach to obtain fMRI data from more practical modalities with high fidelity and downstream applicability.

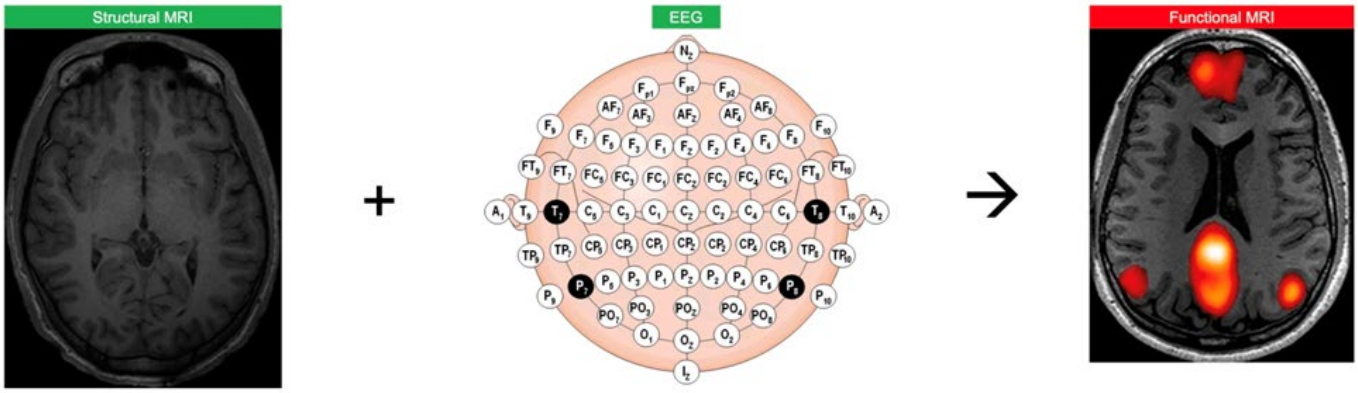


Figure 1: High-level conceptualization of MRI + EEG to fMRI pipeline.

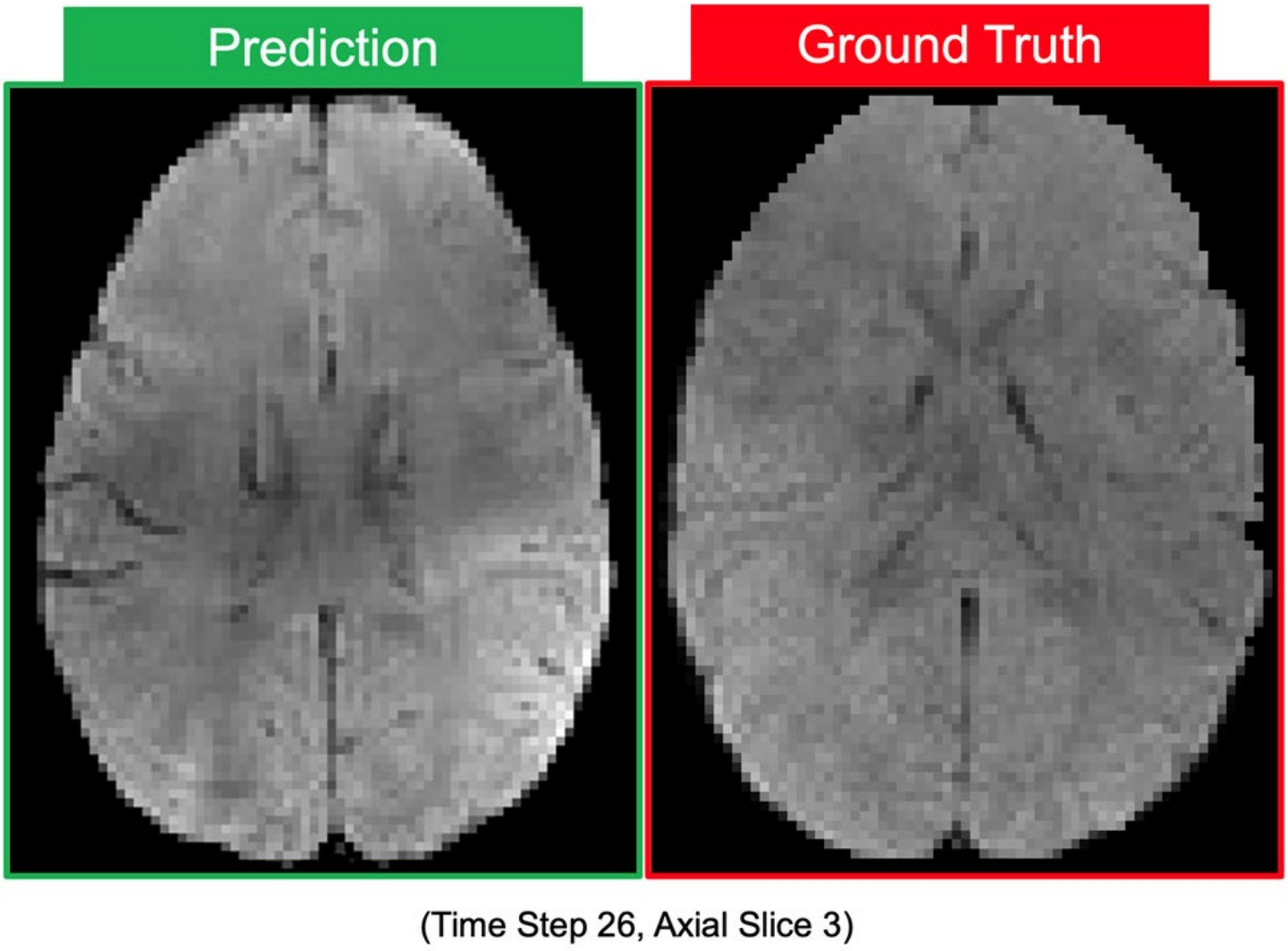


Figure 2: Predicted fMRI and its ground truth from randomly selected subject, time step, plane, and slice.

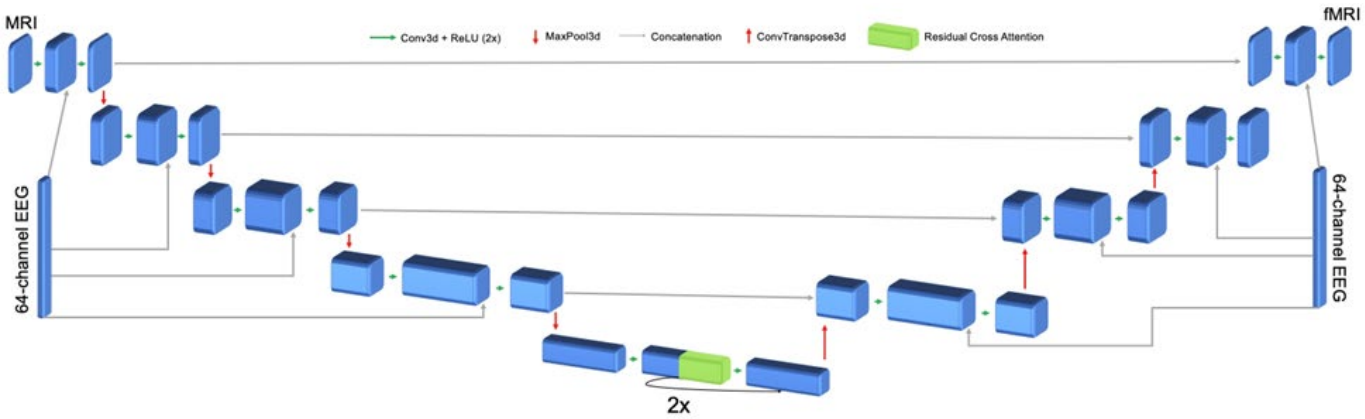


Figure 3: Depiction of the U-Net architecture with cross attention trained to generate fMRI volumes from MRI and EEG inputs.

Keywords
Functional MRI; EEG; Generative AI; Multimodal

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Utilizing Generative AI to Recognize Racial Disparities in Imaging Registries: A Step Toward Model Explainability

Bardia Khosravi, MD, MPH, MHPE, Postdoctoral Research Fellow, Mayo Clinic AI Laboratory;
Pouria Rouzrokh, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Hillary W. Garner, MD;
Doris E. Wenger, MD; Michael J. Taunton, MD; Cody C. Wyles, MD

Introduction

Disparities in population and medical databases can introduce bias, particularly when training deep learning (DL) models, as these biases may be propagated without detection.

Hypothesis

This study aimed to use generative DL technology to understand radiographic differences based on race among patients undergoing total hip arthroplasty (THA), the second most common surgery in the United States.

Methods

We used 480,000 radiographs from our well-characterized hip and pelvis radiographic registry, which includes images from more than 15,000 patients who underwent THA between 1997 and 2021. Patient demographic information (age, sex, BMI, and race) was extracted from the electronic health record. We used a generative diffusion probabilistic model (DDPM) to create synthetic radiographs conditioned on the given patient's demographic information and image characteristics. We generated 60 videos of pelvic radiographs of a Caucasian patient being transformed into those of an African American patient while controlling for age, sex, and BMI (Figure 1). Two MSK radiologists and one orthopedic surgeon carefully examined these interpolation videos and were asked to characterize systematic differences in the images. We used Gwet's AC1 (GAC) to measure the agreement among the readers, with values >0.60 designating substantial agreement.

Results

Expert evaluators identified 5 characteristics that were systematically and consistently different among the 30 pairs of images. The group identified that African American patients when compared to Caucasian patients demonstrated 1) increased acetabular protrusio (GAC: 0.88), 2) higher degree of osteoarthritis (GAC: 0.83), 3) elongated pelvis and more elliptical obturator foramen (GAC: 0.80), suggesting higher lumbar lordosis, 4) decreased femoral neck-shaft angle (GAC: 0.75), and 5) increased femoral metaphyseal cortical thickness (GAC: 0.56). P-value in all instances were < 0.001.

Conclusion

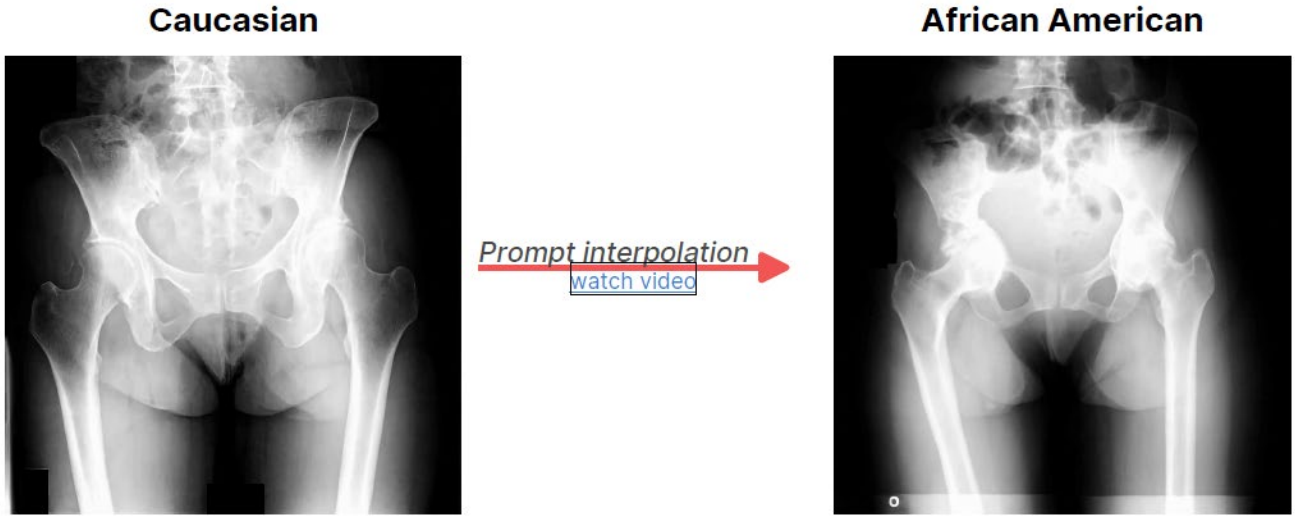
Our results show that DDPMs, when properly conditioned, can be used to visualize racial disparities among patients. It is important to note that these findings do not necessarily reflect anthropomorphic differences and only show differences among patients who have undergone total hip arthroplasty at a tertiary center. Further studies

should be conducted to systematically evaluate these findings at other institutions.

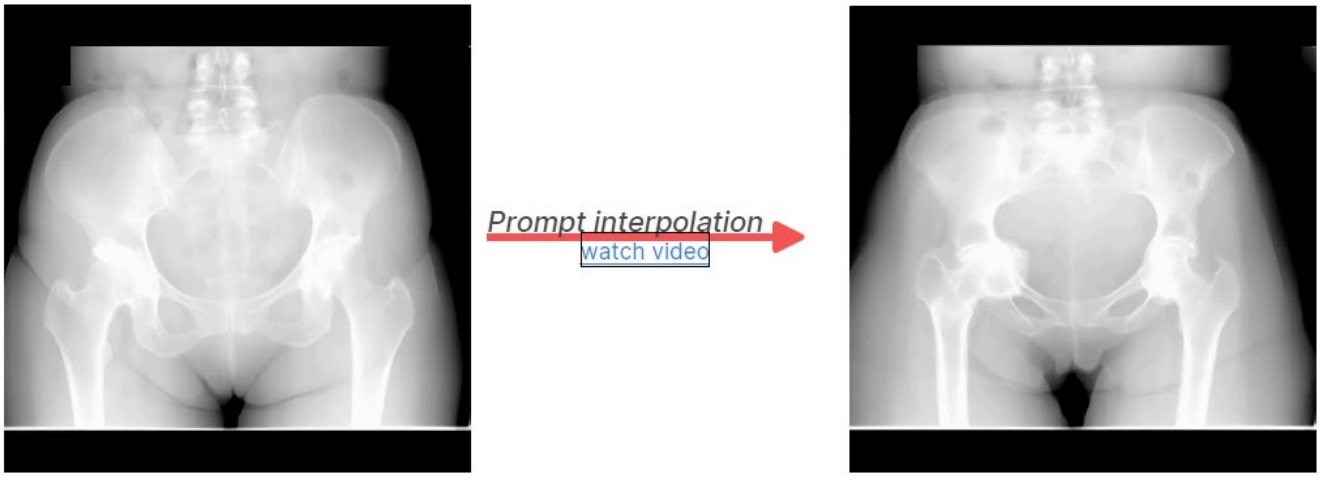
Statement of Impact

DL models are highly effective at extracting information from medical images. It has been shown that these models can detect a patient's race from a single radiograph, but it is not clear what features the models are using. In this work, we demonstrate an explainability method that can shed light on some of the differences perceived by these models.

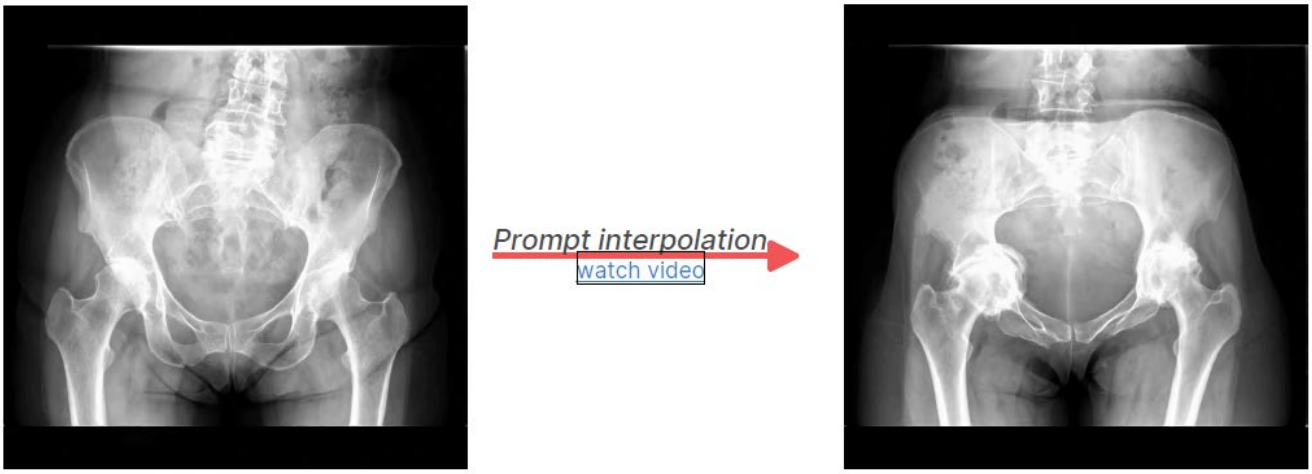
Examples of the generated images. We used embedding interpolation with the same initial noise, to have the closest pelvis radiograph to a given image but from a different race.



Watch Prompt Video: https://drive.google.com/file/d/1ezOh1WFRT88agOfLK_A03MgycbDdRjMM/view



Watch Prompt Video: https://drive.google.com/file/d/1Bt9GcA2pOKCFxTf02j30mTwr3SxQoNdx/view



All images are created with the following prompt and the specific {race} label :

"AP radiograph of a {race} Female in their 50s with a BMI of 25-30, with both joints present and no prosthesis."

Watch Video: https://drive.google.com/file/d/1mvzAfv-YyYYmIZbIScv5jm6YuRNZENFM/view

Keywords

Generative AI; Bias; Diffusion Models; MSK Radiology

Scientific Abstract Presentations
Toolkits and Machine Learning Algorithms

Date:
MON, OCT 2

Time:
10:30 AM – 12:00 PM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

A Comprehensive Guide to Preparing Medical Imaging Data for AI: A SIIM Survey
+ Sanaz Vahdati, MD, Postdoctoral Research Fellow, Mayo Clinic
+ Bardia Khosravi, MD, MPH, MHPE; Elham Mahmoudi, MD, MPH; Pouria Rouzrokh, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Aylin Tahmasebi, MD; Katherine Andriole, PhD, FSIIM; Peter Chang, MD; Keyvan Farahani, PhD; Mona G. Flores, MD; Judy W. Gichoya, MD, MS; Sina Houshmand; MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Care to ExplAIn? Differential Impacts of Explanation Types on Physician Trust in AI
+ Drew Prinster, PhD Candidate, Johns Hopkins University
+ Amama Mahmood; Suchi Saria, PhD; Jean Jeudy, MD; Cheng T. Lin, MD; Chien-Ming Huang, PhD; Paul H. Yi, MD, MS

Evaluating the Utility of Self-Configuring Capsule Networks for Brain Image Segmentation
+ Durga Sritharan, Postgraduate Associate, Yale School of Medicine
+ Sanjay Aneja, MD; Arman Avesta, MD, PhD; Rahul D'Souza; Mariam Aboian, MD, PhD; MingDe Lin, PhD

Exploring Interpretation Maps as a Path to Discover Radiogenomics Biomarkers: A Call for Rethinking
+ Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic Rochester
+ Mana Moassefi, MD; Gian Marco Conte, MD, PhD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

From Isolation to Collaboration: Harmonizing Heterogeneous Medical Imaging Datasets with Partial Annotations
+ Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine
+ Adway Kanhere, MS; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Machine Learning for the Prediction of Osteopenia/osteoporosis Using the Bone Attenuation of Multiple Osseous Sites from Chest Computed Tomography
+ Ronnie Sebro, MD, PhD, Musculoskeletal Radiologist, Mayo Clinic Florida
+ Cynthia De la Garza-Ramos, MD

A Comprehensive Guide to Preparing Medical Imaging Data for AI: A SIIM Survey

Sanaz Vahdati, MD, Postdoctoral Research Fellow, Mayo Clinic; Bardia Khosravi, MD, MPH, MHPE; Elham Mahmoudi, MD, MPH; Pouria Rouzrokh, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Aylin Tahmasebi, MD; Katherine Andriole, PhD, FSIIM; Peter Chang, MD; Keyvan Farahani, PhD; Mona G. Flores, MD; Judy W. Gichoya, MD, MS; Sina Houshmand; MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction

The increasing rate of Artificial Intelligence(AI) model development to address numerous clinical challenges has escalated the need to prepare high-quality clinical imaging data. Optimal data preparation is of paramount importance since it leads to the development of standard, reproducible AI models and alleviates biases. The ideal tool should assist developers and researchers in preparing the data in the fastest and most well-curated manner. To achieve this, one should be aware of existing tools with core features that provide most of the tasks at hand. Becoming familiar with existing tools can be beneficial for not only selecting the best tool for the assigned task (e.g., detection, segmentation, or classification) but also can point out the possible limitations the user might face if starting to work with the wrong tool.

Hypothesis

The current study was designed to collect the most used tools for data curation prior to using them to train AI models.

Methods

A questionnaire including the tool's name, description, and core features was prepared and distributed among 54 active members of the Society of Imaging Informatics(SIIM) from 26 medical informatics centers. Duplicates, general answers not describing a specific tool, and institutional custom (not publicly available)-based applications were excluded, resulting in a total of 30 tools. The tools were investigated based on the core features, cloud features, input data, data curation including de-identification functions and data annotation, workflow, federated learning support, and data storage. We categorize different steps of data preparation with a comprehensive collected list of tools that are designed for each task.

Results

Tools with their extracted core features are described in the attached tables for data identification, data curation, and data annotation and labeling. Image normalization, data conversion, input and output data, labeling and segmentation, and auto-segmentation application are some of the main tasks depicted in our findings. It is noted to mention that data curation is defined as the process of selecting and organizing data from the time it is acquired to the point it is ready for use by AI; Thus, data de-identification and annotation can be considered as part of data curation. As depicted, several tools may provide more than one of the aforementioned tasks.

Conclusion

We propose a comprehensive guideline of stages for data curation for AI applications and present lists of tools in each step for optimizing the decision-making.

Statement of Impact

Providing a list of tools for different steps of medical imaging data preparation for AI model development.

Table1. Tools used for data de-identification.

Deidentifier tool	Cloud-based	Viewer	Input/Output	Data storage
Anonymizer (by John Perry - RSNA)			DCM/DCM+ spreadsheet	✓
Flywheel	✓	✓	DCM/DCM	✓
Glendor PHI Sanitizer	Both		Image folder + text data	
CTP (Clinical Trial Processor)		✓	DCM/DCM	
Horos	Both	✓	DCM+database in txt files with comments or notes on single studies	

Table2. Tools used for data curation with their specific core features.

Tool	De-identification	Viewer	Cloud-based	Input/Output	Conversion	Normalization
MD.ai	✓	✓	✓	Dcm/Dcm, CSV, JSON	✓	
MONAI Label			Both	DCM, NiFTI/DCM, NiFTI	✓	✓
3D slicer	✓	✓	Both	DCM, NiFTI/DCM, NiFTI	✓	✓
ImageJ/FIJI	✓	✓		DCM, NiFTI/DCM, NiFTI	✓	✓
The Medical Imaging Interaction Toolkit (MITK)		✓		DCM, NiFTI/DCM, NiFTI	✓	✓
dicom2nifti				DCM/NiFTI	✓	
DCM2niix	✓			DCM/NiFTI	✓	
MOOSE				DCM/NiFTI	✓	✓
Niffler	✓			DCM/NiFTI	✓	
MANGO		✓		DCM, NiFTI/NiFTI	✓	✓
ITK-SNAP		✓		DCM, NiFTI/NiFTI	✓	✓

Table3. Tools used for data annotation with their specific core features.

Tool	Cloud-base	Input/ Output	Labeling	Segmentation	Active learning	Object detection	3D rendering	Co-registration	classification
MD.ai	✓	Dcm/Dc m, CSV, JSON		✓+ Auto segmentation		✓			✓
MONAI Label	Both	DCM, NiFTI/DC M, NiFTI	✓	✓+ Auto segmentation	✓				✓
3D slicer	Both	DCM, NiFTI/DC M, NiFTI	✓	✓+ Auto segmentation	✓		✓	✓	

ITK-SNAP		DCM, NiFTI/Nif TI		✓+ Auto segmentation			✓	✓	
ImageJ		DCM, NiFTI/DCM, NiFTI	✓	✓			✓	✓	
ImageJ/FIJI		DCM, NiFTI/DCM, NiFTI	✓	✓		✓	✓	✓	✓
The Medical Imaging Interaction Toolkit (MITK)		DCM, NiFTI/DCM, NiFTI		✓+ Auto segmentation			✓	✓	
Ril-Contour		NiFTI/Nif TI		✓+ Auto segmentation			✓		
NCI Imaging Data Commons (IDC)	✓	DCM, NiFTI/DCM, NiFTI	✓						
MOOSE		DCM/Nif TI (3D)		Auto segmentation					
MANGO		DCM, NiFTI/Nif TI	✓				✓		
Prodigy		DCM, NiFTI/DCM, NiFTI (2D)+ text data			✓	✓			✓
Horos	Both	DCM+ database in text files with comment or notes on single studies	✓						
Markit	✓		✓			✓			✓

Keywords
Artificial Intelligence; Data preparation; Data curation; Toolkit

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Care to ExplAIn? Differential Impacts of Explanation Types on Physician Trust in AI

Drew Prinster, PhD Candidate, Johns Hopkins University; Amama Mahmood; Suchi Saria, PhD; Jean Jeudy, MD; Cheng T. Lin, MD; Chien-Ming Huang, PhD; Paul H. Yi, MD, MS

Introduction
Although AI is increasingly used in clinical practice, few studies have evaluated how different AI explanation methods and other design considerations like AI uncertainty communication, might impact physician diagnostic performance or trust in AI advice. We evaluated how AI explanation types, confidence levels, and correctness would impact physician diagnostic performance for chest x-ray (CXR) diagnosis.

Hypothesis
AI explanation types will impact physician diagnostic performance and trust in AI advice.

Methods
We conducted a prospective randomized experiment of 220 physicians (132 radiologists, 88 internal or emergency medicine physicians) who were asked to evaluate 8 CXR cases (Fig.1). We evaluated how AI explanation types, confidence levels, and advice correctness impact diagnostic performance, confidence in final diagnosis, and perception of AI advice. The between-subjects factor was the AI explanation type: either a “local” explanation that localized key CXR image features with a bounding box, or a “global” explanation that compared a given CXR image to a prototypical example of the CXR in question. AI advice correctness and AI confidence were varied as within-subjects factors. We analyzed the data using appropriate generalized linear mixed-effects (GLME) models. Finally, we evaluated if our key findings can be understood through a ‘novel ‘simple trust” mechanism.

Results
Local AI explanations increased physician diagnostic accuracy compared to global AI explanations when AI advice was correct (Fig.2). Furthermore, AI confidence and task expertise modulated the effect of explanation type on diagnostic accuracy. Physicians used local AI explanations more efficiently (time required for diagnosis) than global AI explanations, though explanation type did not impact physician’s subjective perceptions of AI advice or their confidence in their final diagnosis. Lastly, we identified a potential explanation for our results via a novel heuristic for “simple trust”—which can be roughly understood as reliance without verification—that suggests that physicians tend to more quickly align their diagnosis with local AI explanations than global AI explanations, regardless of AI advice correctness.

Conclusion
AI explanation types impact physician diagnostic performance and trust in AI advice more than the physicians themselves are aware of. AI explanation types substantially impact multiple behavioral measures (diagnostic accuracy, efficiency, and simple trust in AI advice). AI developers and clinicians alike should carefully consider the differential impacts of AI explanation types on diagnostic performance and trust when designing and using AI systems.

Statement of Impact
AI Explanation mechanisms impact how physicians perform in CXR diagnosis and their trust in AI advice, which need to be considered during clinical deployment of AI.

Figure 1: Illustration of experimental setup and procedure: (A) Participating physicians first view the X-ray case without AI advice. (B) Once ready, physicians view AI advice including AI explanation and AI confidence, with design conditions applied here. (C) Physicians then decide whether or how to use AI advice, finalize their diagnosis, provide their confidence, and rate AI advice usefulness. (D) Physicians then repeat for 8 X-ray cases.

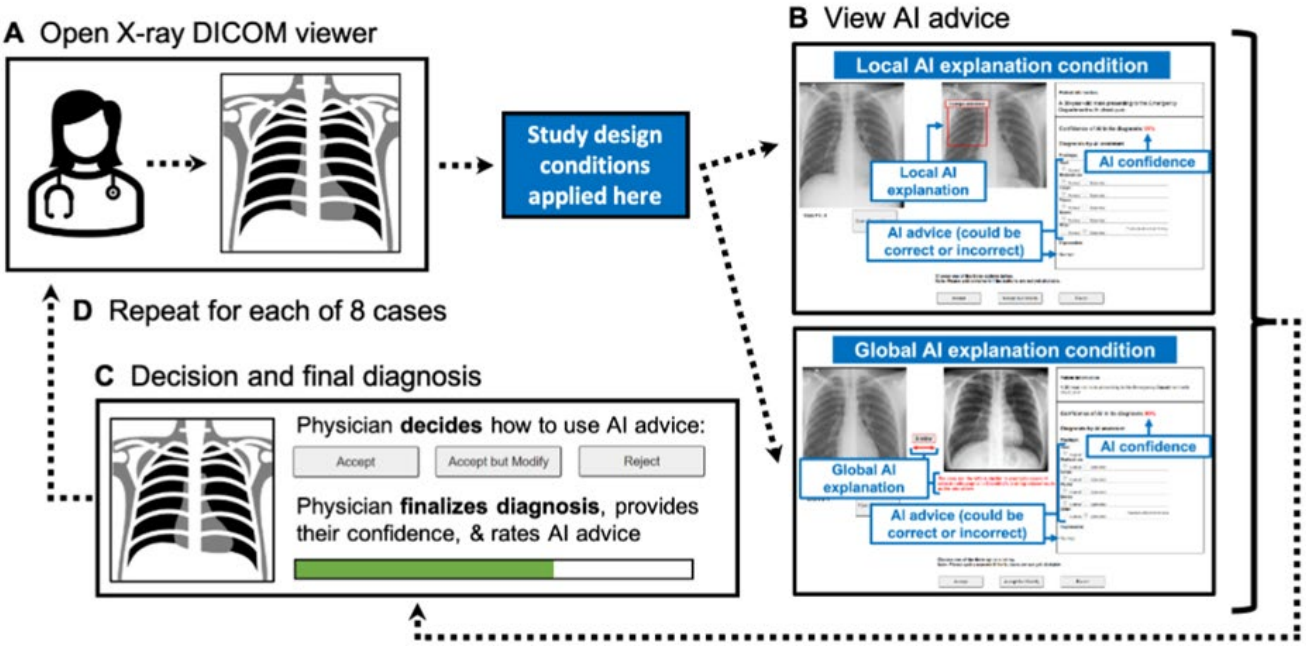
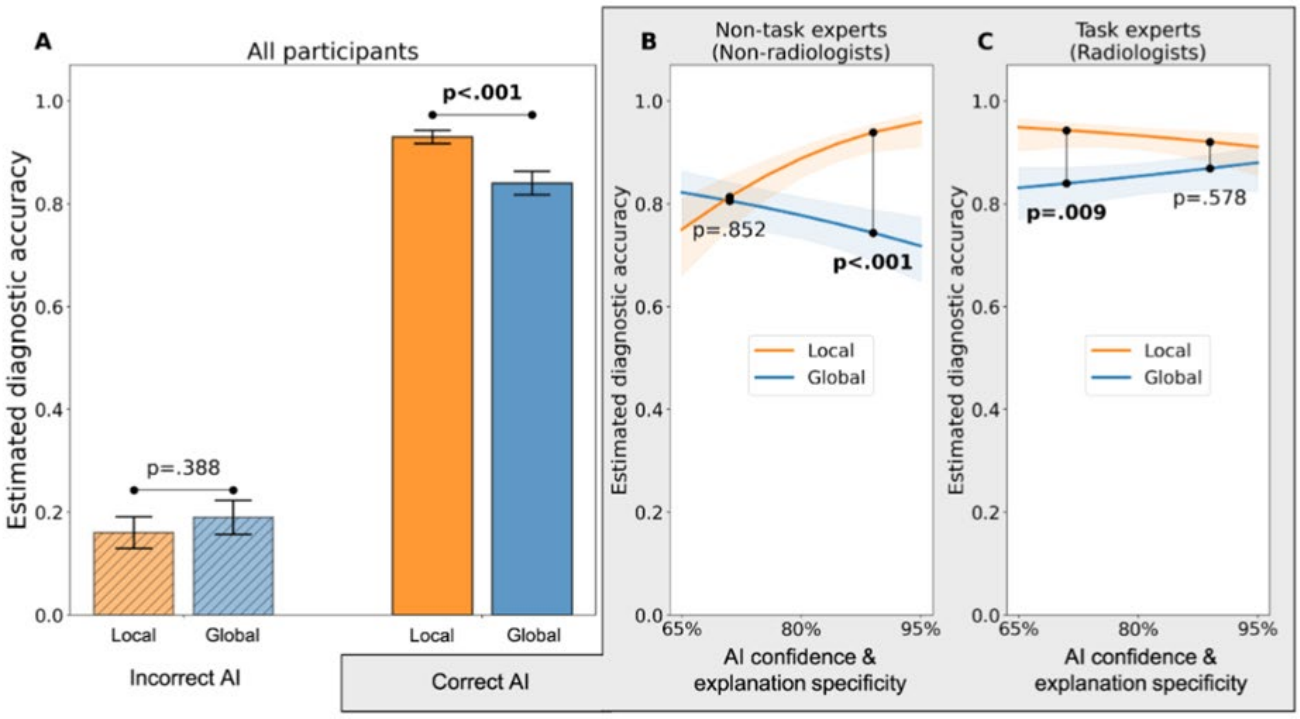


Figure 2: Main results for diagnostic accuracy outcome: Interaction plots for significant interaction effects among experimental variables (addressing primary research questions) for the outcome of marginal-mean estimated diagnostic accuracy. Figure 2A displays the interaction plot for explanation type × AI advice correctness (interaction coefficient β = 1.092, p = .001 [p_adj =]) from the mixed-effects logistic regression model, demonstrating that the impact of AI advice correctness on physician diagnostic accuracy depends on the type of explanation used by the AI: In particular, local AI explanations improve diagnostic accuracy relative to global explanations when AI advice is correct (β = 0.859, p < .001), but we cannot conclude if the explanation type alters the impact of incorrect AI advice on diagnostic accuracy (β = −0.234, p = .388). Only among the correct AI advice condition, figures 2B and 2C display the interaction plots for explanation type × AI advice confidence × physician task expertise (three-way interaction coefficient β = −1.034, p = .012), where radiologists are considered task experts and non-radiologists are non-task experts. In particular, figure 2B illustrates that for non-task experts given correct AI advice, local explanations improve physician diagnostic accuracy relative to global explanations when AI confidence is high (figure 2B, β = 1.598, p < .001), but we do not observe such a difference when AI confidence is low (figure 2B, β = 0.065, p = .852 [p_adj =]). In figure 2C, on the other hand, we see that for task experts given correct AI advice, local explanations improve diagnostic accuracy relative to global explanations when AI confidence is low (figure 2C, β = 1.115, p = .009), but we do not observe such an effect when AI confidence is high (figure 2C, β = 0.578, p > .05).



Keywords
Chest X-ray; Explainability; Human computer interaction; Deep Learning; Trust

Evaluating the Utility of Self-Configuring Capsule Networks for Brain Image Segmentation

Durga Sritharan, Postgraduate Associate, Yale School of Medicine; Sanjay Aneja, MD; Arman Avesta, MD, PhD; Rahul D'Souza; Mariam Aboian, MD, PhD; MingDe Lin, PhD

Introduction

Although a number of deep learning techniques leveraging convolutional neural networks have shown promise for anatomical segmentation, they often require significant amounts of computational memory and training data. Capsule networks represent an alternative and potentially more efficient method for image auto segmentation. We sought to evaluate the utility of self-configuring capsule networks for diagnostic image segmentation.

Hypothesis

We hypothesized that self-configuring capsule networks would be a more computationally efficient method for image segmentation while maintaining high fidelity.

Methods

Using a dataset of 755 MRIs for patients diagnosed with high grade gliomas across multiple facilities within a single healthcare system we trained a self configuring capsule network for tumor identification. Specifically, self-configuring capsule networks were trained to segment tumor enhancing core on the post-T1 contrast sequences. 603 MRIs were used for training, 76 MRIs used for validation, and an additional 76 MRIs were used as a blinded test set. The self-configuring paradigm of the capsule network algorithm included automated adjustments for slice thickness, MRI imaging parameters, and computational resources available for training. Self configuring capsule network performance was compared to traditional convolutional U-NET based auto-segmentation techniques. Dice scores were used to evaluate segmentation performance. Model convergence time, deployment time, and model size (GB) were used to evaluate computational efficiency.

Results

The self-configuring capsule networks showed high fidelity in tumor delineation among gliomas within our dataset. The segmentation accuracy between self-configuring capsule networks was similar to traditional convolutional U-NET based auto-segmentation techniques (89% vs 88%, $p = 0.27$). Self-configuring capsule networks had notably shorter convergence time during training compared to U-NET based models (11 hours vs 38 hours, $p < .001$) and similar deployment time (4 min vs 3.5 min, $p=.15$). Self-configuring capsule networks required significantly less memory when compared to traditional U-NET based segmentation techniques (5 GB vs 31 GB, $p < .001$).

Conclusion

Self configuring capsule networks are a promising computationally efficient method for diagnostic image segmentation tasks with performance that rivals traditional convolutional U-NET based deep learning auto

segmentation techniques. Further studies can help elucidate the potential utility of these novel algorithms within clinical practice.

Statement of Impact

Self-configuring capsule networks are a novel method for image auto segmentation which have the advantage of exquisite computational efficiency compared to alternative deep learning based methods. As the number of algorithms potentially being deployed for image analysis increase, there is an increasing need for more computationally efficient methods which are scalable across all settings.

Keywords

Capsule Networks; Autosegmentation; Deep learning

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Exploring Interpretation Maps as a Path to Discover Radiogenomics Biomarkers: A Call for Rethinking

Shahriar Faghani, MD, Postdoctoral Research Fellow, Mayo Clinic Rochester; Mana Moassefi, MD; Gian Marco Conte, MD, PhD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction

Deep learning (DL) has demonstrated promising results in predicting genetic status based on imaging data. However, these models often lack interpretability. Integrated gradients (IG) is a technique that assigns importance scores to pixels, providing insight into the model's decision-making process and the contribution of each pixel to the output. This study explores the role of IG maps in interpreting the classification of the isocitrate dehydrogenase (IDH) gene and O-6-methylguanine-DNA methyltransferase (MGMT) promoter methylation status in glioblastomas using MRI.

Hypothesis

Interpretation maps can facilitate the discovery of new imaging biomarkers for radiogenomics.

Methods

We analyzed the publicly available The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset, which includes various MRI sequences and corresponding tumor genetic profiles. We trained 3D-Densenet121 models using these sequences independently and in combination to predict IDH mutation and MGMT promoter methylation status. 3D IG maps were utilized to identify the imaging areas that influenced the decision-making process.

Results

The area under the receiver operating curve (AUC) for IDH classification was 0.94, 0.93, and 0.94 for T2, contrast-enhanced T1 (CT1), and T2-CT1 sequences, respectively. The model trained on T2 for MGMT prediction achieved an AUC of 0.65. The model's attention for IDH classification on T2 primarily focused on cerebrospinal fluid, while on CT1, it emphasized contrast-enhanced areas. However, the model primarily highlighted the tumoral area when both CT1 and T2 images were used. Conversely, the model for the MGMT task highlighted all areas without a specific anatomical or functional preference. (Figure 1)

Conclusion

In radiogenomics, the imaging features the model detects, such as IDH classification, depend on the imaging sequence. Using multiparametric MRI directs the model's attention more towards the tumor region. However, no specific imaging areas are emphasized when the model fails to predict the radiogenomics outcome accurately. To draw comprehensive conclusions, employing more interpretation maps, analyzing different combinations of MRI sequences, and considering neuroradiologists' reports on the highlighted areas is necessary. This approach

presents a novel method for identifying imaging biomarkers for radiogenomics through DL.

Statement of Impact

Interpretation maps hold the potential to facilitate the discovery of new imaging biomarkers.

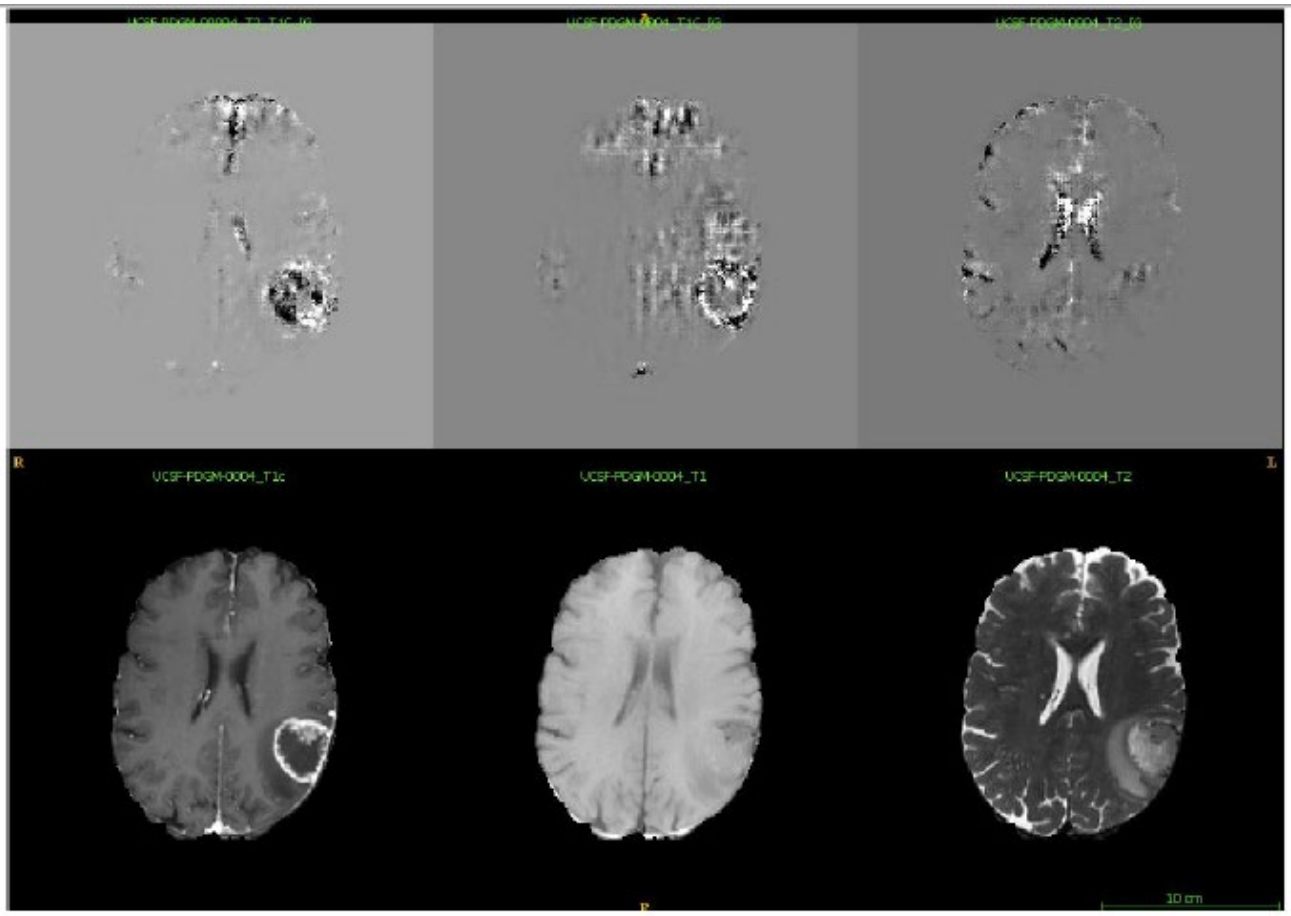


Figure 1 illustrates the brain magnetic resonance imaging (MRI) of a patient with glioma, showcasing T1, T2, and contrast-enhanced T1 (CTE1) images. The top row of the figure displays three distinct integrated gradient maps generated by three different model for predicting the isocitrate dehydrogenase gene status trained on CTE1 and T2 data, CTE1 data alone, and T2 data alone respectively. Notably, all the images within the figure depict the same slice for comparative analysis.

Figure 1 illustrates the brain magnetic resonance imaging (of a patient with glioma, showcasing T 1 T 2 and contrast enhanced T 1 (CTE 1 images The top row of the figure displays three distinct integrated gradient maps generated by three different model for predicting the isocitrate dehydrogenase gene status trained on CTE 1 and T 2 data, CTE 1 data alone, and T 2 data alone respectively Notably, all the images within the figure depict the same slice for comparative analysis

Keywords

Deep learning; Interpretation maps; biomarkers; IDH; MGMT; Glioma

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

From Isolation to Collaboration: Harmonizing Heterogeneous Medical Imaging Datasets with Partial Annotations

Pranav Kulkarni, Bioinformatics Software Engineer, University of Maryland School of Medicine;
Adway Kanhere, MS; Paul H. Yi, MD, MS; Vishwa S. Parekh, PhD

Introduction

Although several large-scale chest x-ray datasets have facilitated the development of deep learning (DL) models, heterogeneity in disease labeling schemes limits their inter-operability [Figure 1]. For example, two datasets with different disease labeling conventions cannot be used to directly train a single DL model. Partial annotations – where datasets have non-overlapping disease labels – prevent use of these datasets in aggregate for DL model training. We developed a collaborative learning framework called surgical aggregation to harmonize imaging datasets with heterogeneous label schemes into a single model even in the presence of partial annotations.

Hypothesis

Surgical aggregation will allow for high-performing DL models trained from datasets with partial annotations that outperform conventional approaches.

Methods

Surgical aggregation is a semi-supervised, model-and-task-agnostic framework that selectively aggregates task-specific knowledge from each participating client to train a global model across all observed labels [Figure 2]. Each client can contribute knowledge for its tasks without interacting with any other client or may choose to import knowledge for different tasks via the global model. We evaluate surgical aggregation's ability to harmonize the NIH CXR14 (n=112,120) and CheXpert (n=224,316) datasets using 70%-10%-20% train-validation-test splits (split at patient level). Each dataset has 14 disease labels, of which 7 are common, for a total of 20 unique labels. We used these datasets to train a 20-label classifier, with external testing on the MIMIC-CXR-JPG dataset (n=377,110). We evaluate the surgical aggregation model using mean area under the ROC curve (AUROC) and compare to models trained using conventional methods (baseline, central aggregation, and federated learning) using bootstrapping and paired t-tests; significance was defined as $p< 0.05$.

Results

On the NIH test set (n=22,330), surgical aggregation performed comparably to the NIH baseline with an AUROC of 0.81 ($p=0.06$), while outperforming other conventional approaches (AUROC of 0.67-0.68; $p< 0.001$) and outperformed all conventional approaches with an AUROC of 0.75 on the CheXpert test set (n=45,208) compared to 0.69-0.74 ($p< 0.001$). Similarly, on the external MIMIC test set, surgical aggregation outperformed all approaches with an AUROC of 0.75 compared to 0.71-0.72 ($p< 0.001$) [Figure 3].

Conclusion

Surgical aggregation allows for harmonization of datasets with heterogeneous and non-overlapping disease labeling conventions to train high-performing DL models for CXR diagnosis. This method can scale to any medical imaging use case with heterogeneously labeled datasets.

Statement of Impact

As DL-assisted disease characterization becomes a mainstay in radiology, surgical aggregation provides a framework to leverage heterogeneous medical imaging datasets in aggregate to train large-scale clinically-useful models.

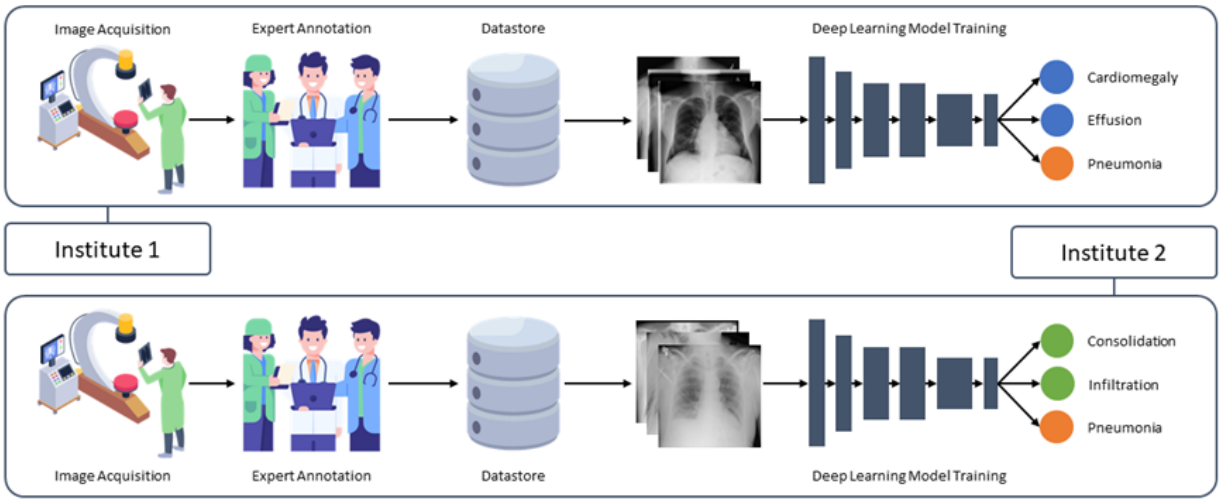


Figure 1. Siloed approach of developing and training models separately on large-scale medical imaging datasets. In this illustration, both institutes curate chest x-ray datasets to train deep learning models on similar tasks. However, due to data and label heterogeneity and patient data privacy, harmonizing and leveraging knowledge from both datasets is difficult.

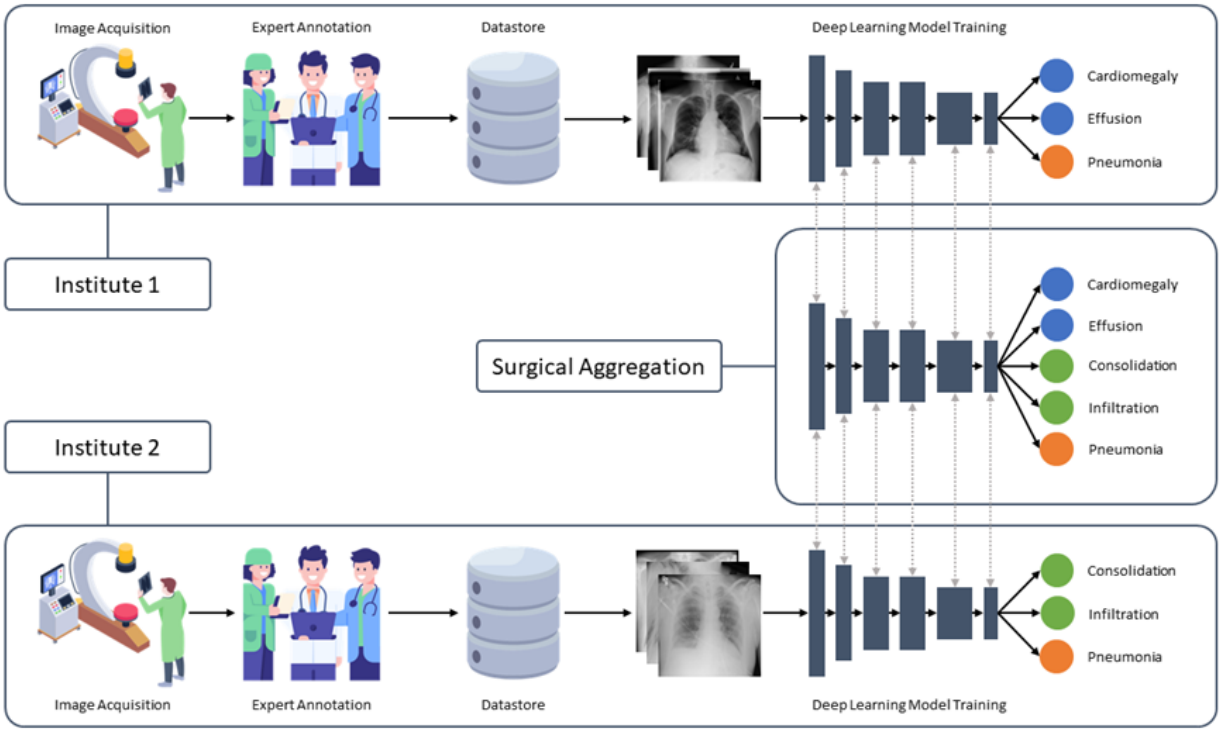


Figure 2. An overview of the surgical aggregation framework. Due to inherent differences in image acquisition, annotation, and curation, large-scale medical imaging datasets are heterogeneous and focus on similar but different disease annotations. Surgical aggregation harmonizes and aggregates knowledge from these heterogeneous datasets into a global deep learning model.

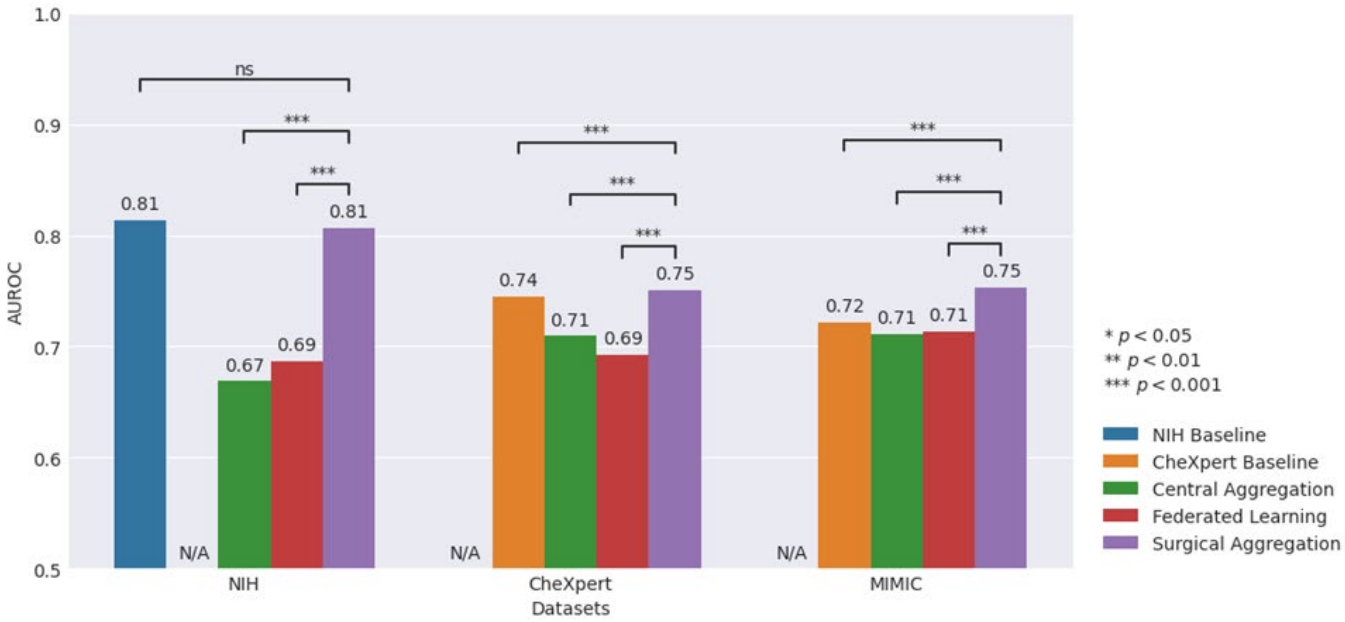


Figure 3. Comparison between the mean AUROC score metrics of surgical aggregation, central aggregation, federated learning, and baseline models on held-out NIH and CheXpert test sets and external MIMIC test set across all 20 disease labels.

Keywords
Data heterogeneity; Data harmonization; Federated learning; Collaborative learning; Chest x-ray; Classification

Machine Learning for the Prediction of Osteopenia/osteoporosis Using the Bone Attenuation of Multiple Osseous Sites from Chest Computed Tomography

Ronnie Sebro, MD, PhD, Professor of Radiology, Mayo Clinic Florida; Cynthia De la Garza-Ramos, MD

Introduction
The attenuation of the lumbar and thoracic spine trabecular bone from CT scans of the abdomen and pelvis have shown utility in predicting BMD measurements. Prior studies have used 2D assessments of the vertebral body trabecular CT attenuation measured on a single slice, however, we hypothesized that volumetric 3D measurement of the vertebral body trabecular CT attenuation would provide a better estimate of the patient's BMD than 2D measurements because the entire trabecular vertebral body is assessed rather than a sample.

Hypothesis
To use machine learning and the 3D CT attenuation of all bones visible on chest CT scans to predict osteopenia/osteoporosis.

Methods
We retrospectively evaluated 364 patients with chest CT and Dual-energy X-ray absorptiometry (DXA) scans within 6 months of each other between 01/01/2015-08/01/2021. Volumetric segmentation of the ribs, thoracic vertebrae, sternum, and clavicle was performed using 3D Slicer to obtain the mean CT attenuation of each bone. The study sample was randomly split into training/validation (80%, n=291) and test (20%, n=73) datasets. Univariate analyses were used to identify the optimal CT attenuation thresholds to diagnose osteopenia/osteoporosis. We used penalized multivariable logistic regression models including Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Ridge regression, and Support Vector Machines (SVM) with radial basis functions (RBF) to predict osteopenia/osteoporosis and compared these results to the CT attenuation threshold at T12.

Results
There were positive correlations between the CT attenuation between all bones ($r>0.6$, $P< 0.001$ for all). There were positive correlations between CT attenuation of the bones and the L1-L4 BMD T-score, total hip T-score, and femoral neck T-scores ($r>0.4$, $P< 0.001$ for all). A CT attenuation threshold of 170.2 Hounsfield units (HU) at T12 had an AUC of 0.702, while a threshold of 192.1 HU at T4 had an AUC of 0.757. The SVM with RBF had the highest AUC (AUC=0.864) and was better than the LASSO (P=0.011), Elastic Net (P=0.011), Ridge regression (P=0.011) but was not better than using the CT attenuation at T12 (P=0.060).

Conclusion
The CT attenuation of the ribs, thoracic vertebra, sternum, and clavicle can be used individually and collectively to predict BMD and to predict osteopenia/osteoporosis.

Statement of Impact
Although the CT attenuation of T12 has been historically used to screen for osteopenia/osteoporosis, we found that a T4 CT attenuation threshold of 192.1 HU had a higher AUC than a T12 threshold of 170.2 HU.

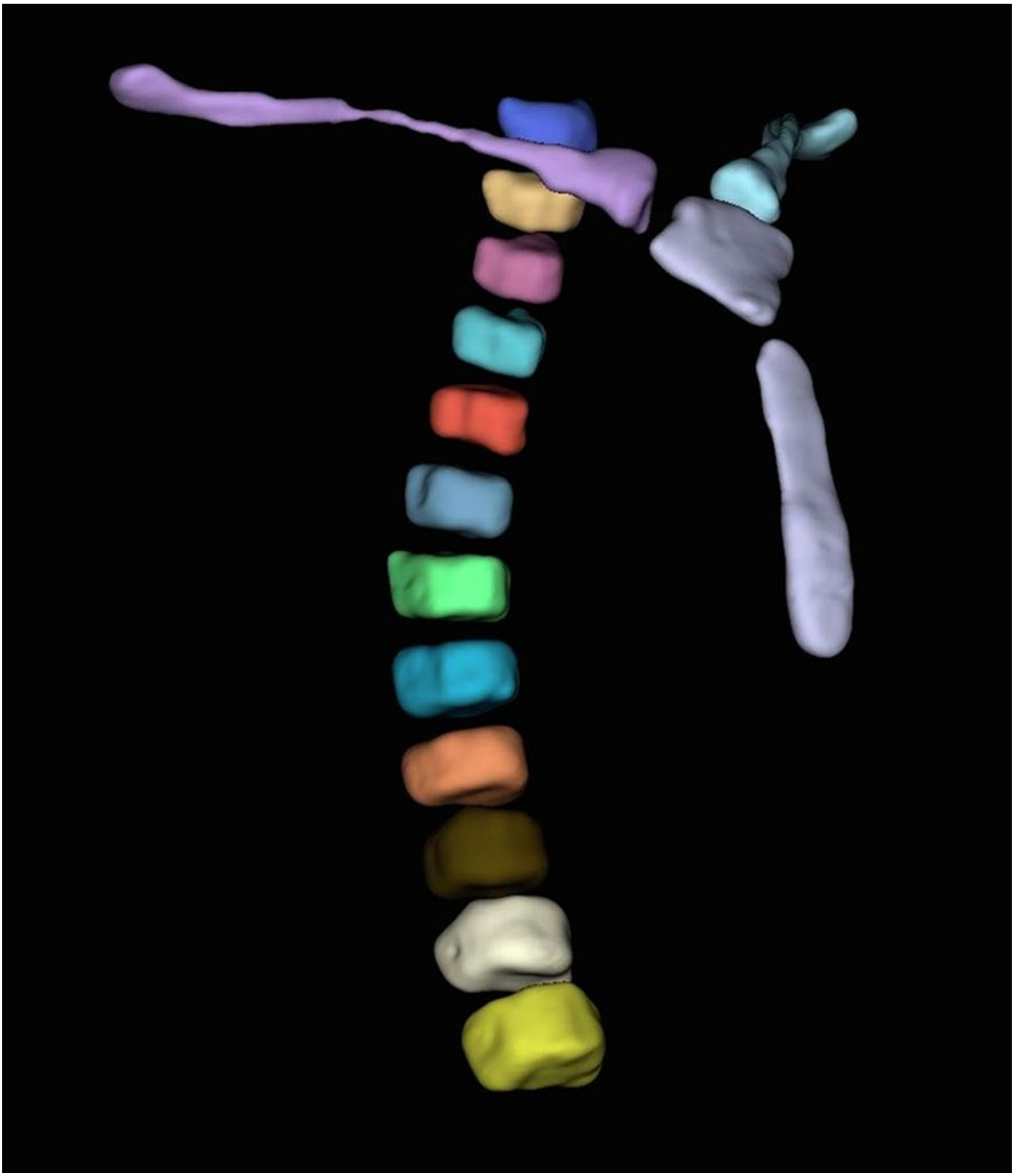


Fig. 1. Three-dimensional volumetric segmentation of the sternum, clavicles, and thoracic vertebrae (T1-T12).

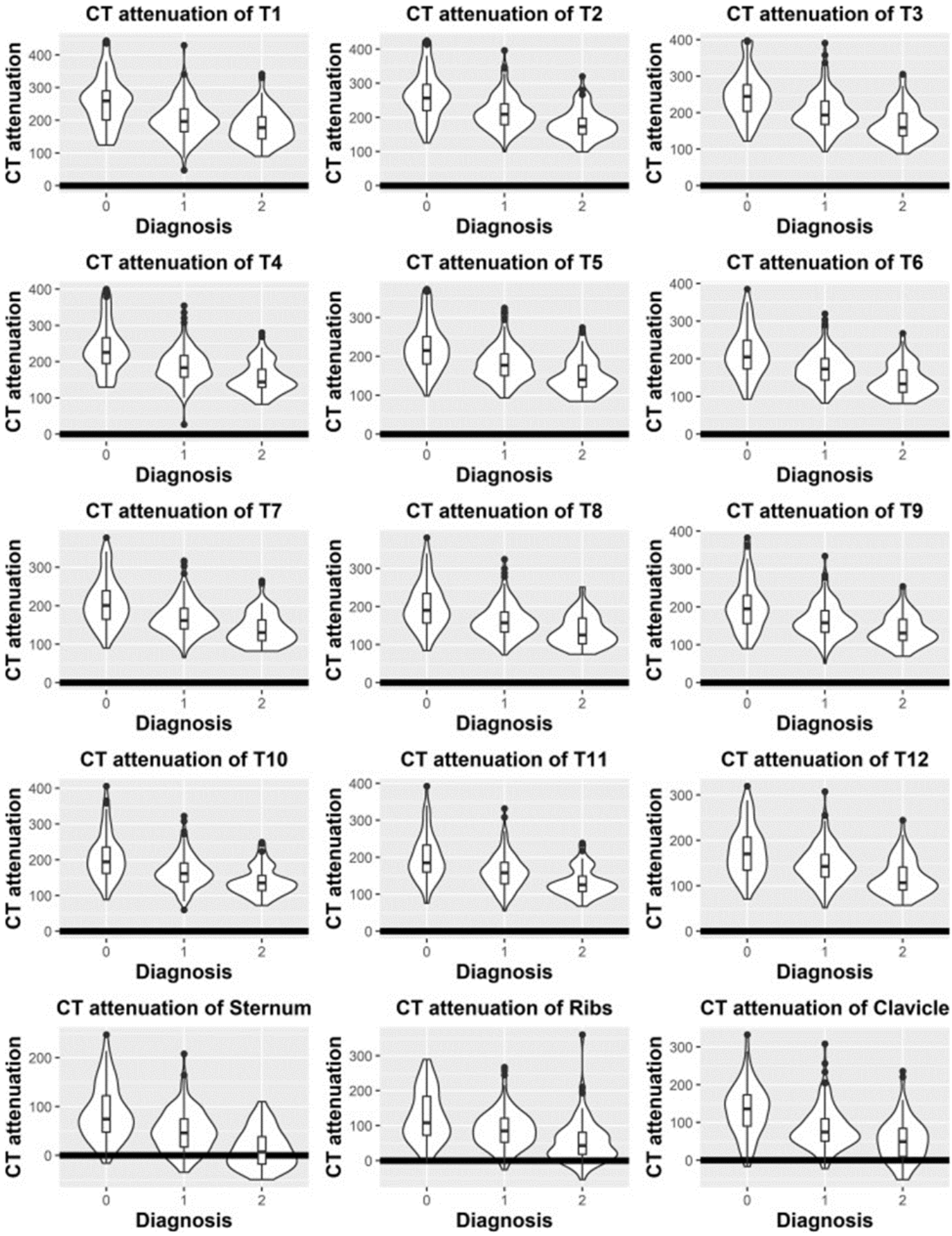

CT attenuation of T1
CT attenuation of T2
CT attenuation of T3
CT attenuation of T4
CT attenuation of T5
CT attenuation of T6
CT attenuation of T7
CT attenuation of T8
CT attenuation of T9
CT attenuation of T10
CT attenuation of T11
CT attenuation of T12
CT attenuation of Sternum
CT attenuation of Ribs
CT attenuation of Clavicle
CT attenuation
Diagnosis
0
1
2

Fig. 2. Violin plots of the mean CT attenuation of each bone by WHO diagnosis. 0 = Normal BMD. 1 = Osteopenia. 2 = Osteoporosis.

Keywords
thoracic spine; computed tomography; osteoporosis; osteopenia; bone mineral density

Scientific Abstract Presentations
Clinical Applications

Date:
MON, OCT 2

Time:
1:00 PM – 2:15 PM ET

Location:
Turner Auditorium

Continuing Education:
ASRT-RT | CAMPEP-MPCEC | SIIM IIP-CIIP

Coarse Race and Ethnicity Labels Mask Granular Underdiagnosis Disparities in Deep Learning Models for Chest X-Ray Diagnosis
+ Preetham Bachina, Medical Student, Johns Hopkins School of Medicine
+ Sean P. Garin, Adway Kanhere, MSE; Pranav Kulkarni, Vishwa S. Parekh, PhD; Jeremias Sulam, PhD; Paul H. Yi, MD, MS

Evolutionary Strategies of AI to Study Language Dominance on Functional MRI
+ Joseph N. Stember, MD, PhD, Assistant Attending, Radiology, Memorial Sloan Kettering Cancer Center
+ Katherine Dishner; Mehrnaz Jenabi, MS; Holodny Andrei, MD; Kyung Peck, PhD; Hrithwik Shalu, MS

HAnging Protocol Prediction by Image Identification (HAPPII)
+ Jennifer Oettinger, Undergraduate Research Intern, Mayo Clinic
+ Alex Bratt, MD; Daniel Blezek, PhD

Pediatric-specific nnU-Net and DeepMedic Methods for Autosegmentation of Brain Tumors: A Comparison Study
+ Arastoo Vossough, PhD, MD, CIIP, Associate Professor, Children's Hospital of Philadelphia, University of Pennsylvania
+ Nastaran Khalili, MD, MPH; Anahita Fathi Kazerooni, PhD; Debanjan Haldar, MD; Karthik Viswanathan, MS; Ariana Familiar, PhD; Sina Bagheri, MD; Hannah Anderson, MS; Aria Mahtabfar, MD; Wenxin Tu; Ibraheem Salman Sheikh, MD; Phillip Storm, MD; Adam Resnick, PhD; Christos Davatzikos, PhD; Jeffrey B. Ware, MD; Ali Nabavizadeh, MD

THA-AID: A Trustworthy Deep Learning Tool for Total Hip Arthroplasty Automatic Implant Detection with Uncertainty and Outlier Quantification
+ Pouria Rouzrokh, MD, MPH, MHPE, Research Associate, Radiology, Mayo Clinic
+ John Mickley, MD; Bardia Khosravi, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Bradley Erickson, MD, PhD, CIIP, FSIIM; Michael Taunton, MD; Cody Wyles, MD

SIIM CMIMI23
OCTOBER 1-2
Johns Hopkins | Baltimore, MD
Conference on Machine Intelligence in Medical Imaging

Coarse Race and Ethnicity Labels Mask Granular Underdiagnosis Disparities in Deep Learning Models for Chest X-Ray Diagnosis

Preetham Bachina, Medical Student, Johns Hopkins School of Medicine; Sean P. Garin, Adway Kanhere, MSE; Pranav Kulkarni, Vishwa S. Parekh, PhD; Jeremias Sulam, PhD; Paul H. Yi, MD, MS

Introduction
Deep learning (DL) models for medical imaging diagnosis can exhibit demographic biases. However, coarse race and ethnicity labels, such as “Asian,” are frequently used to categorize a diverse array of ethnic subgroups, such as “Indian”, “Korean” and “Chinese,” which may conceal non-trivial medical differences . We evaluated if coarse race and ethnicity labels could conceal granular ethnic disparities in underdiagnosis rates in DL models for chest radiograph (CXR) diagnosis.

Hypothesis
DL models will have granular ethnicity biases hidden by coarse race/ethnicity labels.

Methods
We trained DL classifiers (DenseNet, ImageNet-pretrained) to diagnose 14 disease labels using two CXR datasets: MIMIC-CXR ([N=377,095] and CheXpert [N=224,316] each divided into 80/10/10% train/validation/test splits. For each dataset, we developed five models using five random seeds. For testing, we used MIMIC-CXR because it contains coarse and granular race and/or ethnicity labels (self-reported); MIMIC-CXR-trained models were tested on the internal hold-out test set (N=25,888) and CheXpert-trained models were tested on the entire MIMIC-CXR dataset (N=270,136). We calculated mean area under the ROC curve (mAUC) across all disease labels and each set of models; comparisons were performed using DeLong’s test. We used underdiagnosis rate (false positive rate [FPR] for ‘no finding’ label) as our primary bias metric for measuring potential delayed access to care. FPR was calculated using a threshold defined by maximum test set F1-score. Underdiagnosis rates were calculated as the mean of five models with 95% confidence intervals.

Results
Both MIMIC-CXR and CheXpert-trained models had state-of-the-art (SOTA) performance with mAUCs of 0.828±0.001 and 0.803±0.002, respectively (Table 1). Using coarse labels, both models had significantly higher underdiagnosis rates in ‘Black’ (25±0.4%, 34±0.2% respectively) and ‘Hispanic/Latino’ patients (27±0.4%, 33±0.4% respectively) compared to ‘White’ patients (20±0.2%, 28±0.2% respectively; P< 0.005, all) (Figure 1). Sub-analyses by granular labels revealed variations in underdiagnosis rates for all coarse labels and all models that frequently exceeded variation between coarse group labels (Figure 1). For example, for the ‘Hispanic’ label, underdiagnosis rates ranged from 0±0.0% (Honduran) to 96±1.4% (Colombian) and for the ‘White’ label, they ranged from 1±0.2% (Brazilian) to 39±2.1% (Eastern European) in the MIMIC-CXR models.

Conclusion
We reproduced underdiagnosis biases of SOTA CXR DL classification models favoring ‘White’ over ‘Black’ and ‘Hispanic’ patients. However, these coarse labels concealed significant disparities between granular ethnicity groups, which may underestimate their negative impact.

Statement of Impact
Algorithmic biases should be measured using granular ethnicity labels instead of coarse ones, lest these promising technologies inadvertently mask hidden disparities.

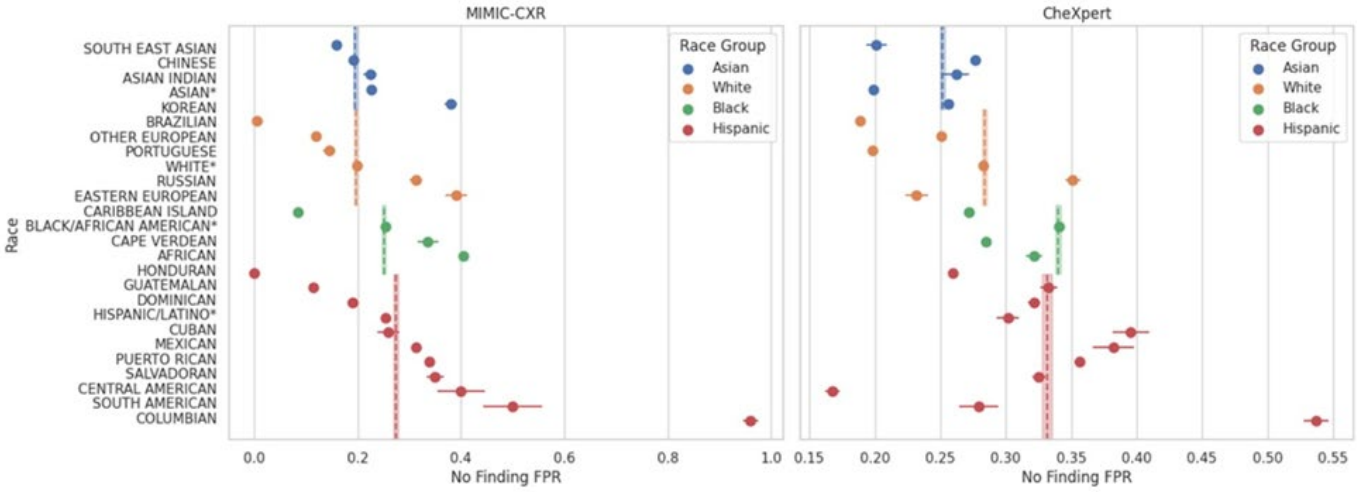
Table 1: State of the Art Performance (SOTA) for models trained on MIMIC and CheXpert.

Table 1: State of the Art Performance (SOTA) for models trained on MIMIC and CheXpert.

Dataset	Our Method (mean AUC ± 95% CI)	*Seyyed-Kalantari et al. (Nature Medicine 2021)* (mean AUC ± 95% CI)	*Cohen et al. (arXiv 2020)* (mean AUC; 95% CI not reported by authors)
MIMIC-CXR(4)	0.828±0.001	0.834±0.001	0.83
CheXpert(5)	0.803±0.002	0.805±0.001	0.8

Abbreviation: CI = confidence interval. Comparison of our mean AUC over all labels on internal test sets with well-established SOTA models from previous works.

Figure 1: Granular Underdiagnosis Rates ('No Finding' False Positive Rate [FPR]) for models trained on MIMIC-CXR and CheXpert. Points show averages and solid lines show 95% confidence intervals (CIs) for granular group FPR across 5 models. Dashed lines and shaded regions show averages & 95% CIs for coarse groups. Granular groups labeled with an asterisk * are the patients who only reported a coarse race.



Keywords
Deep Learning; Bias; Fairness; Chest X-Ray

Evolutionary Strategies of AI to Study Language Dominance on Functional MRI

Joseph N. Stember, MD, PhD, Assistant Attending, Radiology, Memorial Sloan Kettering Cancer Center; Katherine Dishner; Mehrnaz Jenabi, MS; Holodny Andrei, MD; Kyung Peck, PhD; Hrithwik Shalu, MS

Introduction

Preoperative planning often involves using functional magnetic resonance imaging (fMRI) to identify critical language functions in relation to brain tumors. Language comprehension and production are primarily associated with specific brain areas, namely Wernicke's and Broca's areas. Determining laterality, especially for individuals who exhibit right-dominance or co-dominance, is crucial in this process. While task-based fMRI (tb-fMRI) is the current standard for determining language dominance, it may not provide optimal or diagnostic results for certain cancer patients, including those with dementia, hearing loss, or language barriers. We propose employing deep learning techniques on resting-state fMRI (rs-fMRI) to address this challenge of determining language laterality.

Hypothesis

We hypothesized that a convolutional neural network (CNN) trained with evolutionary strategies could learn to predict language laterality from rs-fMRI adjacency matrices.

Methods

We retrospectively obtained a total of 60 rs-fMRI exams. All patients also had tb-fMRI, providing gold-standard language localization. The breakdown was: 30 left-dominant, 9 right-dominant, and 21 co-dominant. We divided into training and testing sets as follows: training set consisted of 15 left-dominant cases, 5 right-dominant, and 10 co-dominant. The testing set was also 15 left-dominant, 4 right-dominant, and 11 co-dominant. All rs-fMRI BOLD signal spectra were post-processed into adjacency matrices. The adjacency matrices provided the input to a relatively simple CNN consisting of 4 convolutional layers, followed by 3 fully connected layers, with 2-node output for the two classes of left-dominant versus non-left-dominant (either right-dominant or co-dominant). We trained the CNN with deep neuroevolution (DNE), an evolutionary-based approach that has shown promise for small training sets, such as ours.

Results

Training for 30,000 evolutionary generations, we obtained a maximum testing set accuracy of 80%. The model did not overfit by defaulting to the majority class (left-dominant); the prediction accuracy was 12/15 for left-dominant, 3/4 for right-dominant, and 9/11 for co-dominant.

Conclusion

Deep learning with evolutionary strategies can predict language laterality based on rs-fMRI with reasonable accuracy.

Statement of Impact

Deep learning can extract valuable language laterality results for patients who cannot perform tb-fMRI.

Keywords

Evolutionary strategies; Deep neuroevolution; function MRI; fMRI; language laterality

HAnging Protocol Prediction by Image Identification (HAPPII)

Jennifer Oettinger, Undergraduate Research Intern, Mayo Clinic; Alex Bratt, MD; Daniel Blezek, PhD

Background/Problem to be solved

To create efficient radiologist workflows and reduce errors, hanging protocols can be devised to dictate how to arrange radiology images for viewing. However, hanging protocols are tedious to create and maintain. For example in highly complex modalities such as cardiac MRI, heterogeneity in scanner equipment, imaging indication, image acquisition protocol, and scan personnel, can generate thousands of different permutations of series descriptions; our institution's cardiovascular division acquires series with over 6,000 unique series descriptions. To manage these series typically strict logic rules based on series description are used to determine series placement, but slight changes in scanner, technologist, or procedure cause series to be misclassified by these rules and render the hanging protocol ineffective. In the past, machine learning attempts have directly learned hanging positions from series, but these methods resulted in low accuracy (25-33%) and are inflexible with respect to changes in hanging protocols.

Intervention(s)

We propose a new method that uses machine learning to classify each series based on multiple DICOM tags into one of 69 semantic labels, which is then used instead of the series description to create a more robust hanging protocol. Semantic labels were manually chosen by an experienced cardiovascular radiologist with four years of post-fellowship experienced. Due to the high degree of class imbalance (Figure 1), the largest class having 3,648 examples compared to the smallest class with 3, the model uses a two-stage classifier. First, a random forest classifier divides the series into one of 10 super-classes (e.g. cine, delayed enhancement, localizer, etc.), then superclass-specific classifiers assign a specific semantic label and confidence score to each series.

Outcome

On a held-out test dataset of 10 cases (314 series) the model identified series with 98% accuracy by superclass and 88% accuracy by fine-grained semantic label, achieving an overall Matthew's correlation coefficient of 0.88 and AUC of 0.89 (Figure 2).

Conclusion

By excluding the original series description from the DICOM tags the classifier considers, the model is unaffected by noise in series description. Our model also allows for more flexibility than previous artificial intelligence methods as it does not learn a final hanging protocol directly, rather learning how to classify each image into one of the semantic labels, which allows users to change the final hanging protocol without retraining the model.

Statement of Impact

By incorporating this type of model into clinical practice, hanging protocols may be more effective and robust, reducing inconvenience and error caused by previous methods.

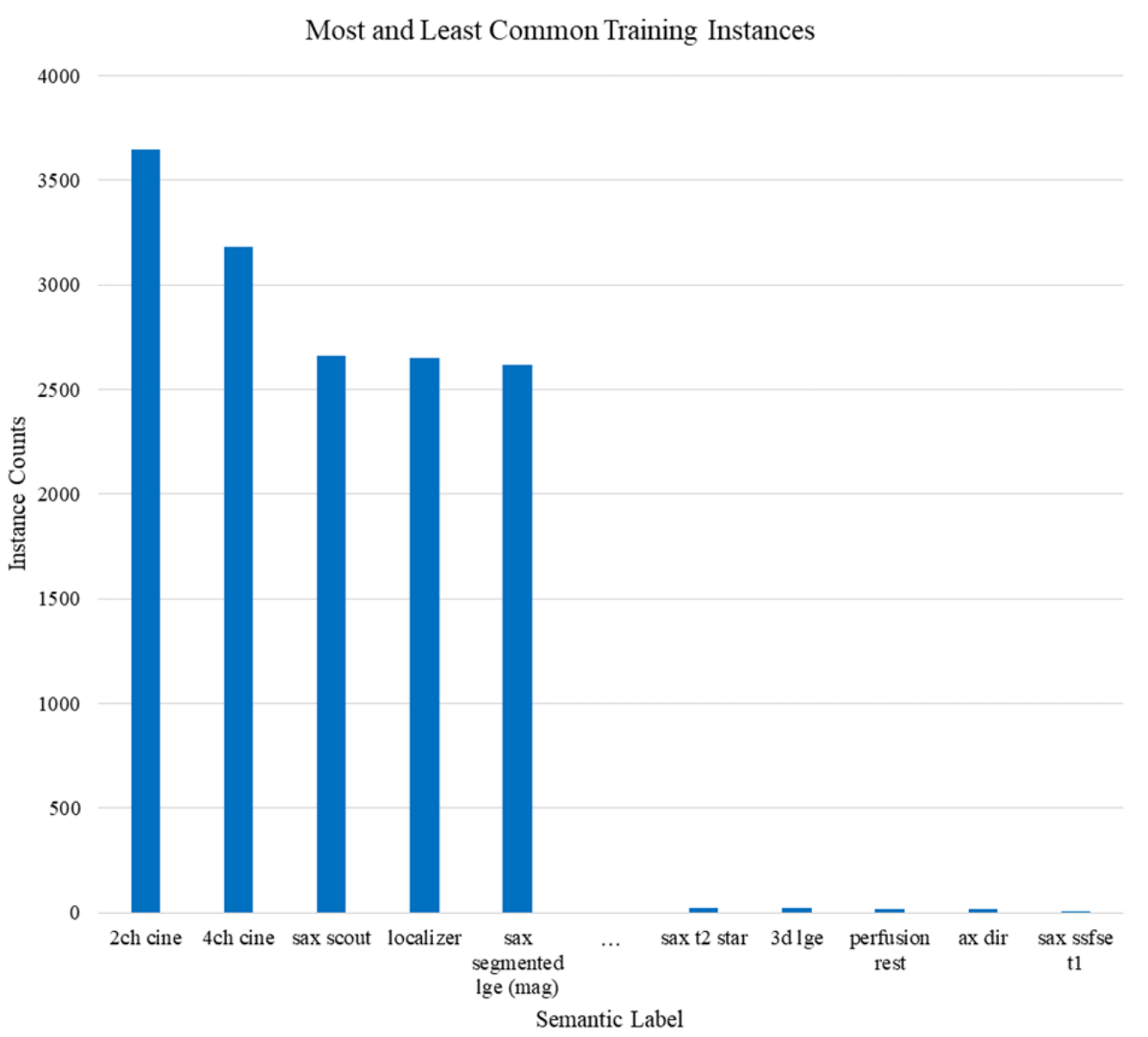


Figure 1: Histogram of the Five Most and Five Least Common Series Descriptions in Training Dataset

Figure 2: ROC curve where true positive corresponds to a correct prediction with confidence greater than threshold, false positive corresponds to an incorrect prediction with confidence greater than threshold, true negative corresponds to an incorrect prediction with confidence less than threshold, and false negative corresponds to a correct prediction with confidence less than threshold.

Keywords
Artificial intelligence; multi-class classification; hanging protocol; image display

Pediatric-specific nnU-Net and DeepMedic Methods for Autosegmentation of Brain Tumors: A Comparison Study

Arastoo Vossough, PhD, MD, CIIP, Associate Professor, Children's Hospital of Philadelphia, University of Pennsylvania; Nastaran Khalili, MD, MPH; Anahita Fathi Kazerooni, PhD; Debanjan Haldar, MD; Karthik Viswanathan, MS; Ariana Familiar, PhD; Sina Bagheri, MD; Hannah Anderson, MS; Aria Mahtabfar, MD; Wenxin Tu; Ibraheem Salman Sheikh, MD; Phillip Storm, MD; Adam Resnick, PhD; Christos Davatzikos, PhD; Jeffrey B. Ware, MD; Ali Nabavizadeh, MD

Introduction

Accurate segmentation of pediatric brain tumors can play an important role in radiomic analysis, treatment planning, and accurate quantitative follow-up. Tumor types, subregions, and follow-up guidelines distinctly differ between adult and pediatric brain tumors. There is a paucity of autosegmentation methods that target pediatric-specific tumor subregions and that are validated across a wide spectrum of pediatric tumor types. Here, we train and compare the newer nnU-Net architecture to DeepMedic, a previously established segmentation platform for automated segmentation of pediatric brain tumors.

Hypothesis

A trained nnU_Net segmentation architecture will be superior for pediatric whole tumor and subregion segmentation compared to DeepMedic.

Methods

Pre-operative multi-parametric scans (T1w, T1w-CE, T2, and T2-FLAIR) of 339 pediatric patients (n=293 internal and n=46 external) with a variety of supratentorial and infratentorial tumors were manually segmented to identify four subregions: enhancing tumor (ET), non-enhancing tumor (NET), cystic components (CC), and peritumoral edema (ED). The internal cohort was split into training (n=233) and test (n=60) subsets for nnU-Net (nnU_NET has built-in validation), and training (n=186), validation (n=47), and test (n=60) subsets and for DeepMedic. Correlations between automated and expert tumor subregion percentage volumes were also calculated.

Results

Dice similarity score for the internal test sets for nnU-Net and DeepMedic (respectively) was: 0.94±0.10 vs 0.88±0.17 for whole tumor (p=0.0001), 0.85±0.33 vs 0.75±0.33 for ET (p=0.0361) and 0.85±0.19 vs 0.73±0.21 for union of all non-enhancing components (NET+CC+ED) (p=0.00004). In the external test set, the Dice score was: 0.90±0.06 vs 0.86±0.20 for whole tumor (p=0.00004), 0.84±0.30 vs 0.74±0.35 for ET (p=0.0216), and 0.80±0.21 vs 0.64±0.25 for union of all non-enhancing components (p=0.0003), respectively. nnU-Net also had higher accuracy for tumor subregions: Dice score for cystic component 0.79-±-0.37 vs 0.43-±-0.37(p=0.0007) and non-enhancing component: 0.80-±-32 vs 0.52-±-0.28 (p=0.0002). The lowest sensitivities and highest 95% Hausdorff distances were for edema in both test sets. Pearson correlation coefficient of volumetric measurements of automated tumor

subregion segmentations vs expert segmentations was 0.93, 0.94, 0.78, and 0.94 (nnU_Net) and 0.90, 0.73, 0.71, and 0.33 (DeepMedic) for ET, NET, CC, and ED regions, respectively.

Conclusion

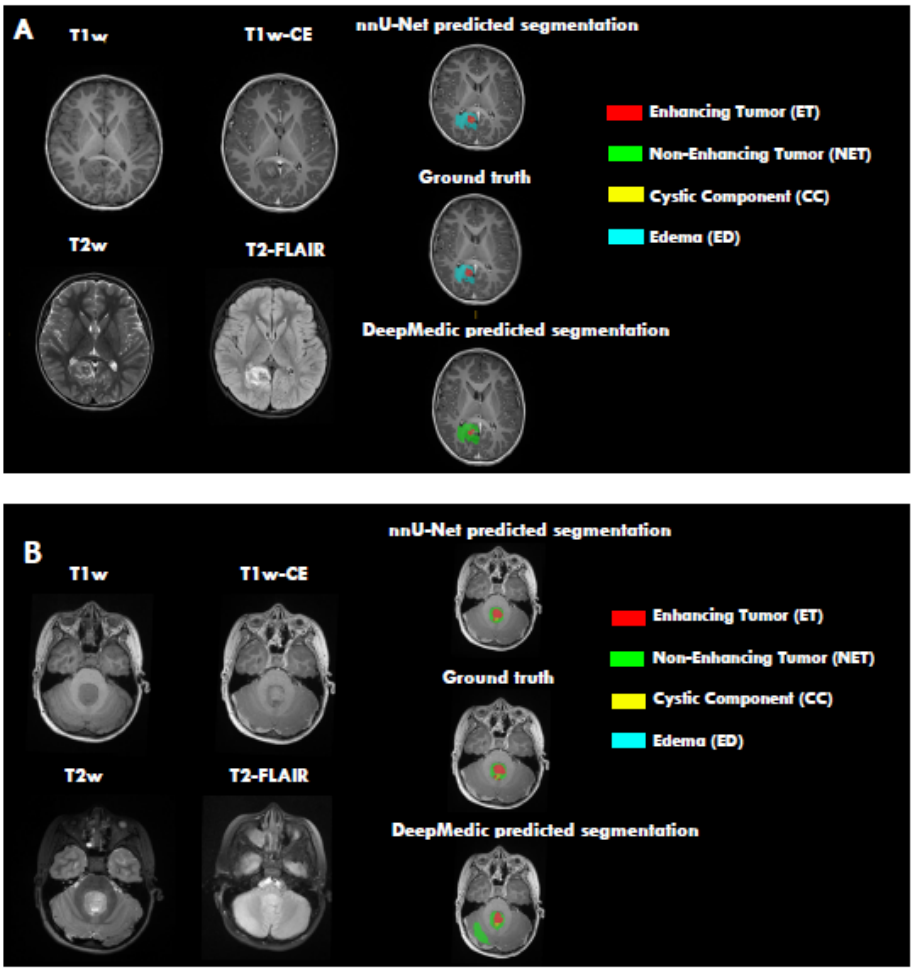
Our newly trained nnU-Net model demonstrated high performance and was superior to the previously established DeepMedic architecture on segmentation of whole tumor and tumor subregions in a cohort of various pediatric brain tumors.

Statement of Impact

This trained nnU-Net platform can be used for rather reliable segmentation of a wide variety of pediatric brain tumors.

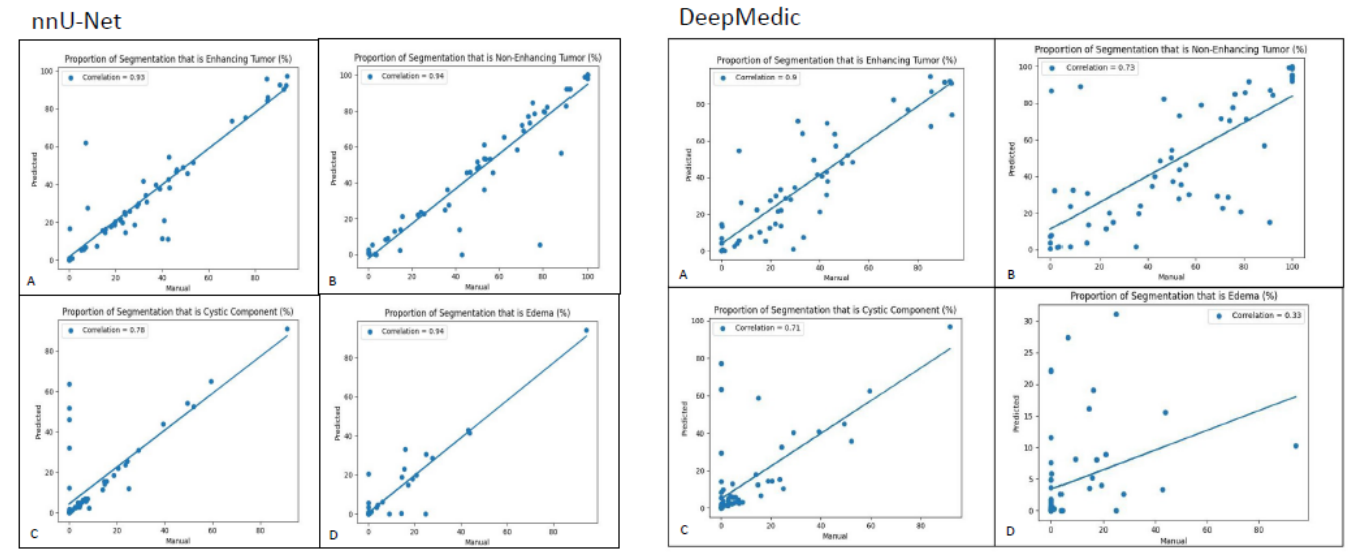
(A): This image shows good performance of both models in outlining the whole tumor region; however, nnU-Net had a better performance in prediction of subregions. (Whole tumor (WT): ET + NET + ED + CC). nnU-Net: WT dice: 0.96; ET dice: 0.88; NET dice:0.67; CC dice:0; ED dice:0.96; NET+ CC + ED dice: 0.96 DeepMedic: WT dice: 0.92; ET dice: 0.59; NET dice: 0.0045; CC dice:0; ED dice:0.20; NET+CC+ED dice:0.91 (B): This image shows superior performance of nnU-Net for automated segmentation of the whole tumor region; however, both models had a good performance in terms of predicting the enhancing tumor component. nnU-Net: WT dice: 0.94; ET dice: 0.86; NET dice:0.84; CC dice:0.61; ED dice:1; NET + CC+ ED dice: 0.86 DeepMedic: WT dice: 0.45; ET dice: 0.75; NET dice: 0.35; CC dice: 0.05; ED dice: 0; NET+CC+ED dice: 0.33

Figure 1



Correlations between predicted and automated tumor subregion percentage volumes for internal test set. Scatter plot of results from the comparison of the agreement between volumes of tumor subregions predicted by nnU-Net and DeepMedic compared to ground truth: (A) proportion of tumor that is enhancing; (B) proportion of tumor that is non-enhancing; (C) proportion of tumor that is cystic (not a VASARI feature); (D) proportion of tumor that is edema.

Figure 2



Keywords

Pediatric; brain tumors; segmentation

THA-AID: A Trustworthy Deep Learning Tool for Total Hip Arthroplasty Automatic Implant Detection with Uncertainty and Outlier Quantification

Pouria Rouzrokh, MD, MPH, MHPE, Research Associate, Radiology, Mayo Clinic; John Mickley, MD; Bardia Khosravi, MD, MPH, MHPE; Shahriar Faghani, MD; Mana Moassefi, MD; Bradley Erickson, MD, PhD, CIIP, FSIIM; Michael Taunton, MD; Cody Wyles, MD

Introduction
Confident deployment of deep learning (DL) models in clinical settings necessitates trustworthiness, which includes the ability to explain predictions, report uncertainty levels, and identify outlier input data (out-of-distribution). Detecting total hip arthroplasty (THA) implants on plain radiographs is crucial for facilitating revision THA, but existing DL models lack trustworthiness for clinical deployment.

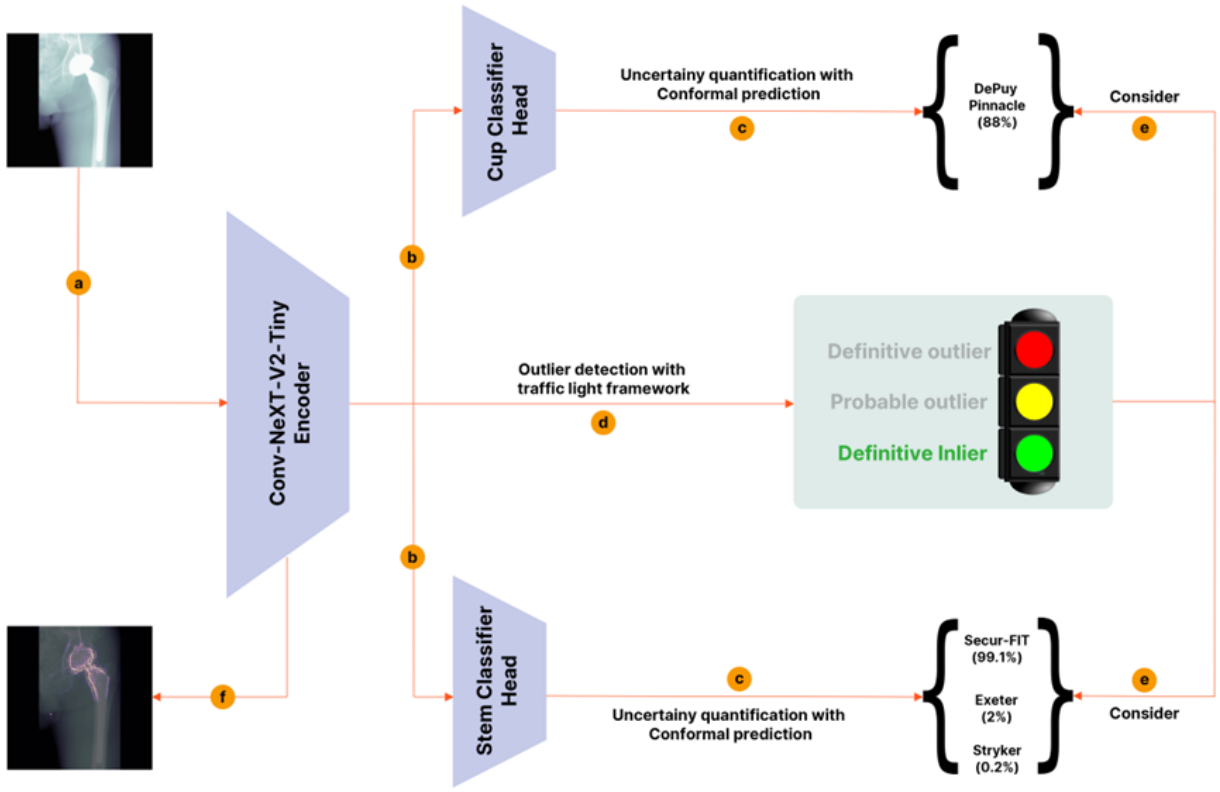
Hypothesis
In this work, we present THA-AID, a trustworthy DL tool for automated implant detection on plain radiographs that not only outperforms all previous models in terms of prediction power, but also is explainable, quantifies the uncertainty of its predictions, and detects out-of-distribution data in inference.

Methods
We developed THA-AID, a 5-fold cross-validation DL classifier to recognize 28 implants in input hip radiographs. THA-AID was trained on 244,248 AP, lateral, and oblique images obtained from 13,375 THA patients (2000-2022). Our trustworthiness pipeline consisted of 1) integrated gradient maps for model explainability, 2) a Mondrian Cross-conformal Predictor (MCCP) for uncertainty quantification that enabled the model to predict more than one label on challenging radiographs, and 3) a framework to detect data outliers, warning users if input image features or MCCP predictions deviate from 95% of the training data (possible outlier). The latter prevented the model from making predictions for out-of-distribution input data. Our pipeline was thoroughly evaluated on held-out internal and external test sets and six medical and non-medical out-of-domain datasets, including ImageNet, RadImageNet, chest X-ray, knee-radiograph, pre-operative hip, and post-operative hip datasets with unseen implants. We also compare our tool's baseline performance with two board-certified orthopedic surgeons.

Results
Accuracy [F1-scores] were 98.9% [0.99] and 98.0% [0.98], with an MCCP coverage [efficiency] of 99.7% [93.7%] and 97% [69.8%] on the internal and external test sets, respectively. Coverage denoted the fraction of prediction sets with true labels, and efficiency denoted the fraction of prediction sets with a single label. THA-AID achieved 97.5% accuracy compared to the best surgeon accuracy of 87% on 200 test images. It also identified 100% of data from out-of-domain datasets. Eliminating 39.7% of the most outlier external data improved the external set's accuracy, coverage, and efficiency to 1.0.

Conclusion
We developed THA-AID, a clinically trustworthy DL model that surpasses all previously reported models in baseline performance and exemplifies a pipeline for making DL models trustworthy.

Statement of Impact
THA-AID is a strong candidate for clinical deployment in radiologic and orthopedic surgery applications. The trustworthiness pipeline exemplified in THA-AID can be easily applied to existing DL models without altering their architecture or training regime.



THA-AID Pipeline Overview. (a) An input image is introduced into the THA-AID convolutional encoder; (b) The encoded features for the input image are then received by the cup and stem classifier heads; (c) Outputs from the cup and stem classifiers are calibrated by the conformal predictor modules to produce prediction sets. As exemplified in this figure, the prediction set for the cup contains one label (representing certain prediction), while the prediction set for the stem contains three labels (indicating uncertain prediction); (d) The outputs from the THA-AID encoder and classifier heads also feed into the traffic-light framework, which categorizes the input data as green (definitive inlier), yellow (probable outlier), or red (definitive outlier); (e) The user is advised to consider the output of the traffic-light framework when interpreting the results from the cup and stem classifiers; (f) In addition to the prediction sets and traffic-light framework outputs, integrated gradient maps are visualized for the input image, providing further assistance for the user in understanding the model's performance.

The most certain (a) and uncertain (b) conformal stem predictions made by THA-AID on the internal test set. In both scenarios, images are ordered from left to right by decreasing uncertainty. For each prediction, the input radiograph is displayed on top, and the integrated gradient map is displayed on the bottom. On top of each input radiograph, the label set predicted for that radiograph, along with the conformal score assigned to each label (in parentheses), is also displayed. Note how the length of prediction sets are longer for uncertain predictions, with the biggest set predicted for an input radiograph with an almost cropped stem. The integrated gradient maps are also much noisier in uncertain predictions.

Data type	Dataset	Size (number of images)	Definitive Inlier (%)	Possible Outlier (%)	Definitive Outlier (%)
In-domain	Internal test set (cups)	29,523	87.40	10.54	2.06
	Internal test set (stems)	29,523	86.87	10.94	2.19
	External test set	405	9.88	57.03	33.09
	Intenral postoperative hip radiographs with unseen cups in training	449	11.11	51.31	37.58
	Internal postoperative hip radiographs with unseen stems in training	2,129	2.87	18.41	78.72
	External hip radiographs with unseen stems in training	397	0.50	7.06	92.44
Out-of-domain	Preoperative hip radiographs	10,000	0.02	0.08	99.90
	Chest radiographs	5,216	0.00	0.00	100
	Knee radiographs	4,796	0.00	0.00	100
	RadImageNet random sample	20,000	0.00	0.03	99.97
	ImageNet random sample	5,000	0.00	0.02	99.98

Performance of the “traffic-light” framework (our proposed outlier detection framework) for detecting outlier data in six in-domain (=images were postoperative hip radiographs) and five out-of-domain (=images were not postoperative hip radiographs) outlier datasets. For each dataset, the size of the dataset (number of images), along with the percentage of inlier, possible outlier, and outlier data are presented.

Keywords

Total hip arthroplasty; Trustworthy artificial intelligence; Uncertainty quantification; Conformal prediction; Outlier detection; OOD detection