

Innovation in Cancer Informatics Summary Report for RROADMAP: Creation of Metastatic Database from Radiology Reports Using Natural Language Processing

This report summarizes our achievements over the last two years. Our initial objectives were to develop natural language processing (NLP) models for annotation of PET/CT reports for metastatic disease, to document patterns of metastatic spread on imaging at large scale. Over the duration of this grant, our key accomplishments are:

1. Development of BERT based language models trained with VASTA and radiologist annotations

In year 1, we leveraged 20,997 human-annotations by VASTA (a partner organization) to train a BERT NLP model to identify presence or absence of metastatic disease. Prior attempts to use BERT NLP models for extraction of metastatic disease on CT reports showed high accuracy [1]. For this proposal, we used the “Impression” text from PET/CT between 2009-2023, from a total of 226,276 radiology reports.

The VASTA team had labels for disease affecting approximately 400 sites, some of which were too granular for our use (e.g. Left lung lower lobe and right lung middle lobe are categorized as separate site by VASTA, but both were combined into a single ‘lung’ category for the purpose of our study). After we reduced the number of possible sites to 50 categories, we calculated the frequency of disease at each organ. As expected, lung, bone, lymph nodes (LNs), and liver are common sites involved by malignancy, and the remaining decrease rapidly after the top 10 sites, as shown in Table 1.

Table 1: Number of positive reports for each metastasis site (LN = Lymph node).

Site	# of positive reports	% of positive records
Lung	8593	41%
Bone	7185	34%
Thoracic and axillary LN	6533	31%
Abdominopelvic LN	6439	31%
Soft tissues	3496	17%
Liver	3148	15%
Stomach	1801	9%
Head and neck LN	1709	8%
Pleura	1547	7%
Peritoneum	1375	7%
Esophagus	1276	6%
Adrenal	1040	5%
Pancreas	723	3%
Bladder	390	2%

Prostate	368	2%
----------	-----	----

We used 70% of 20,997 VASTA records to train site-specific BERT models to predict the presence or absence of metastases and used 30% of the VASTA data to test those models. Our model was based on the BERT architecture, and we performed extensive hyperparameter tuning using the training and evaluation sets. Additionally, we optimized the classification threshold based on the evaluation set to enhance accuracy. Specifically, after each training epoch, we assessed whether adjusting the classification threshold would improve performance. If an improvement was observed, we modified the threshold accordingly. This iterative approach led to performance gains on the test set as well. Initial results were promising, with high accuracies > 90% for most sites (Table 2).

Table 2. Performance of VASTA-trained (V-BERT) model on three representative sites.

Site	Accuracy	Precision	Recall
Liver	97.0%	89.2%	89.5%
Lung	93.6%	86.6%	81.9%
Adrenal	98.9%	88.0%	90.8%

We next wanted to test the V-BERT NLP model performance against radiologists annotated reports for metastatic disease across multiple sites. We first evaluated the V-BERT NLP Adrenal model against 100 random selected reports, which were annotated for metastatic disease by a radiologist. The performance metrics were: accuracy 99%, precision 50%, and recall 100%. While the accuracy was high, this case highlights the limitation of measuring performance on a site (the adrenal) that has a large class imbalance, with about 5% of reports positive in over 20,000 VASTA records, and only 2% true positive reports in our sample of 100 random cases.

Moving forward, we opted to test the BERT NLP models against more balanced sets of reports. Our V-BERT models were used to create such balanced data sets comprised of approximately 50 positive and 50 negative reports for radiologists to annotate, for each organ of interest (Appendix). The V-BERT models were then retested against these balanced radiologist annotations, with noticeable decrease in performance (Table 3).

Table 3. Performance of V-BERT on the VASTA annotation test-set versus and 100 balanced radiologist annotations test-set.

Site	Test set	Accuracy	Precision	Recall
Adrenal	VASTA	98.9%	88.0%	90.8%

Adrenal	Balanced	88%	79.3%	97.7%
Lung	VASTA	93.6%	86.6%	81.9%
Lung	Balanced	68%	39.2%	95.2%

Manual review of the false positive and false negative cases revealed that the poor performance of the V-BERT model arose from the nature of VASTA labels assigned for disease presence, some indicating metastatic disease, some indicating primary malignancies, as well as indeterminate lesions. Thus, the recall rate (or sensitivity) was consistently high, but the precision dropped substantially, especially for the lungs where multiple cases were false positive due to a primary lung cancer being reported in the radiologist impression text.

To address this limitation, we used radiologist annotations to further train the V-BERT models, with a minimum of 500 new annotations per site. This new VASTA and radiologist annotation trained (VR)-BERT model had improved performance, as shown in Table 4.

Table 4. Performance of V-BERT model versus a VASTA and radiologist annotation trained (VR)-BERT, evaluated against a test set of 100 balanced radiologist annotations.

BERT Model	Accuracy	Precision	Recall
V-BERT Adrenal	88%	79.3%	97.7%
VR-BERT Adrenal	92%	88.6%	92.9%
V-BERT Liver	91%	87.7%	96.2%
VR-BERT Liver	92%	89.3%	96.2%
V-BERT Lung	68%	39.2%	95.2%
VR-BERT Lung	85%	58.3%	100%

It was clear that further improvement in performance required further training, and a more strategic approach to labeling reports at scale. This was summarized in a prior ICI progress report, where we sampled ‘marginal’ cases to identify the most challenging reports for the NLP models to predict (Appendix). Marginal cases are those around the threshold between positive and negative reports. However, this approach was time-consuming because of the limitation of obtaining large scale radiologist annotations and it was clear this was not a scalable approach for our goals of labeling all sites.

2. Transition to Large Language Models (LLMs)

In the summer of 2024, we pivoted to the use of Generative AI models, using Chain of Thought and LLaMA 3.1, a state-of-the-art open-source model released by Meta on July 23, 2024. Open-source LLMs are preferable to proprietary LLMs, offering cost-savings and data security [2]. We employed three versions of LLaMA 3.1, with different number of parameters (8B, 70B, 405B). We implemented a Python pipeline that breaks down the classification task into two stages, 1) extraction of sentence(s) relevant to the site of interest for metastatic disease (e.g. liver, lung, or bones) and 2) determination of metastatic disease presence or absence for the output of 1). This approach foregoes the need for human-annotated data, allowing us to use the LLM directly, without additional training. Comparing the three Llama models, it was clear that the smallest version (8B) had substantially lower performance than the 70B and 405B models (Figure 1). While Llama 405B had the best performance overall, it also had computational requirement that was 5 times larger than the 70B model, averaging 20 minutes to process 100 reports. Thus, we opted to use the LLaMA 70B model moving forward. This model demonstrates a performance on par with, and at times surpassing the VR-BERT model (Table 5).

Figure 1. Accuracy of 3 Llama models in predicting metastatic disease measured against balanced radiologist annotations, across 10 organs.

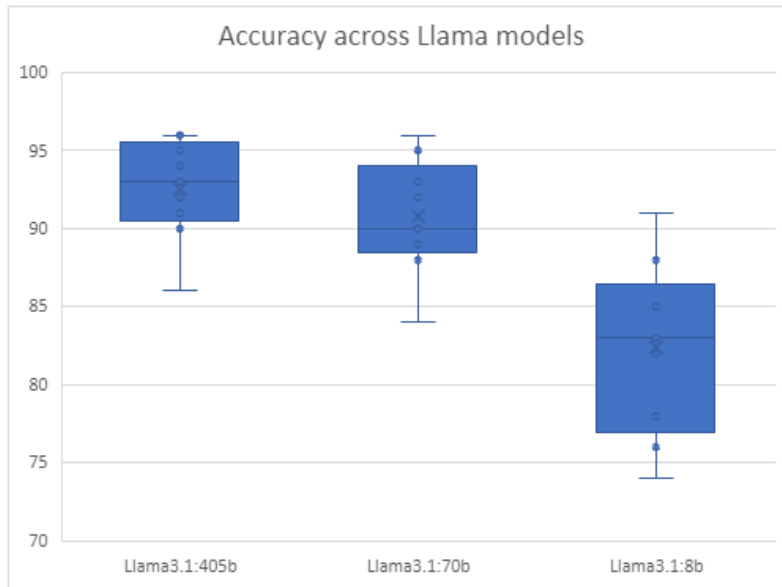


Table 5. Performance of LLaMA 3.1 (70B) against VR-BERT model on 100 balanced radiologist annotations.

Model	Accuracy	Precision	Recall
Llama Adrenal	91%	94.6%	83.3%
VR-BERT Adrenal	92%	88.6%	92.9%

Llama Liver	95%	98.0%	92.3%
VR-BERT Liver	92%	89.3%	96.2%
Llama Lung	97%	87.5%	100.0%
VR-BERT Lung	85%	58.3%	100%

3. Integration of LLaMA and BERT models.

In the final phase of our project, we initially planned to use Llama models to annotate all PET/CT reports across our entire cohort of 84,327 patients and 226,276 reports. However, our calculations showed that it would take approximately six months to annotate these reports with our current workflow. Thus, we explored the possibility of combining the two models, using V-BERT for a majority of predictions, and Llama for V-BERT identified marginal records, i.e. those records most likely to be false positive/false negative cases. V-BERT was retrained at this time, using both CT and PET/CT reports to improve generalizability, using a total of 83,369 VASTA annotated records (20,997 PET/CT reports and 62,372 CT reports).

Our analyses showed an overall performance for V-BERT for 223 marginal records to be quite poor, with Accuracy 67%, Precision 29%, and Recall 48%. The performance of Llama 70B improved for these same records, with the following metrics: Accuracy 86%, Precision 68%, Recall 57%. This supported our strategy for using Llama to substitute marginal predictions by V-BERT. Indeed, when used in combination, the predictions by V-BERT+Llama were comparable to Llama alone for most sites (Table 6), with the benefit of significantly lower computational requirements than for Llama alone. The notable exception is in the lungs, where the VASTA annotations which include primary lung cancer remain an impediment to this approach. Thus, future predictions for lung metastases from PET-CT reports would rely on Llama model exclusively.

Table 6. Performance of Llama, V-BERT, and a combination of the two (V-BERT+Llama) where the V-BERT marginal records are replaced by Llama predictions. AP LN = abdominopelvic lymph nodes.

Site	Model	Accuracy	Precision	Recall
Adrenal	Llama	91%	95%	83%
Adrenal	V-BERT	81%	67%	94%
Adrenal	V-BERT+Llama	87%	77%	92%
Liver	Llama	95%	98%	92%
Liver	V-BERT	91%	88%	96%

Liver	V-BERT+Llama	93%	90%	98%
Lung	Llama	97%	88%	100%
Lung	V-BERT	52%	25%	94%
Lung	V-BERT+Llama	57%	28%	94%
Bone	Llama	94%	100%	87%
Bone	V-BERT	92%	87%	100%
Bone	V-BERT+Llama	92%	88%	98%
AP LN	Llama	89%	92%	73%
AP LN	V-BERT	85%	73%	87%
AP LN	V-BERT+Llama	89%	77%	97%
Peritoneum	Llama	94%	91%	95%
Peritoneum	V-BERT	95%	85%	100%
Peritoneum	V-BERT+Llama	96%	88%	100%

Table 7. Performance improvement from using purely V-BERT versus our combination method (V-BERT+Llama).

Site	Accuracy Change	Precision Change	Recall Change
Adrenal	6.00	10.00	-2.00
Liver	2.00	2.00	2.00
Lung	5.00	3.00	0.00
Bone	0.00	1.00	-2.00
AP LN	4.00	4.00	10.00
Peritoneum	1.00	3.00	0.00

As we near the completion of our project, we plan to propagate our prediction model results on the remaining report impression texts, so we can achieve our initial goal of measuring metastatic disease patterns across all patients at MSKCC with a PET/CT report.

References:

1. Do RKG, etl al. Patterns of Metastatic Disease in Patients with Cancer Derived from Natural Language Processing of Structured CT Radiology Reports over a 10-year Period. *Radiology*. 2021 Oct;301(1):115-122. doi: 10.1148/radiol.2021210043. Epub 2021 Aug 3. PMID: 34342503; PMCID: PMC8474969.

2. Savage CH, Kanhere A, Parekh V, Langlotz CP, Joshi A, Huang H, Doo FX. Open-Source Large Language Models in Radiology: A Review and Tutorial for Practical Research and Clinical Deployment. *Radiology*. 2025 Jan;314(1):e241073. doi: 10.1148/radiol.241073. PMID: 39873598; PMCID: PMC11783163.
3. <https://ai.meta.com/blog/meta-llama-3-1/>

Appendix

1. Creation of Balanced Data Sets

To create more balanced data sets with an even mix of positive and negative cases, the first step is to use a weak classifier. This weak classifier can be a model that has been fine-tuned using a limited number of randomly selected and annotated records. This model will have predictions with low accuracy. However, these predictions can be used to select more balanced records for annotations. In this step, 50% of the positive predictions, and 50% of negative predictions are subsequently selected for annotations.

To select a balanced number of records from each class for annotation, we can use the following mathematical equation:

Let:

- N be the total number of records available for annotation.
- C be the total number of classes.
- N_i be the desired number of records to annotate for class i , where i ranges from 1 to C .

To ensure balance across classes, we aim to annotate an equal number of records for each class. Thus, the desired number of records to annotate for each class is:

$$N_i = N/C$$

where N_i the number of records to annotate for class i .

This equation ensures that each class receives an equal share of the total annotation budget, thereby promoting class balance in the annotated dataset.

2. Labeling of Marginal Records

In this step, we select balanced and challenging records, focusing on those near the marginal

threshold used by the LLM for classification, rather than simply having a 50/50 split of positive and negative predictions. We found that the error rate for these borderline records is typically much higher than for those clearly belonging to one class or the other, and these records are also more difficult for humans to annotate. By including annotations for these challenging records, we observed significant improvements in the model's performance. Incorporating these difficult annotated records into the training data enhances the model's ability to define the boundary between different classes, leading to better overall performance.

Figure 1. Selecting balanced marginal records for annotation

To select balanced marginal records from each class for annotation, we need to identify records that are close to the boundary of class separation as determined by the model.

Let's denote:

- N as total number of records to be selected for annotation in this step

- N_i as the total number of records belonging to class i .

- n as the total number of records to be annotated (assuming the same number for each class).

- M_i as the number of marginal records to be selected from class i .

- x_{ij} as the j -th record from class i .

- $P(x_i)$ as the model's classification confidence for record x_i .

We can define a function $g(x_{ij})$ that measures the rank of record x_{ij} from the decision boundary:

$$g(x_{ij}) = \text{distance}(P(x_i)) \text{ from classification border for class } j$$

Now, to select marginal records from each class, we can sort the records within each class based on $g(Px_{ij})$ in ascending order and select the top M_i records from each class. The equations for selecting balanced marginal records from each class for annotation can be summarized as:

$$M_i = \text{select } (N/C) \text{ records from each class } j \text{ based on } g(x_{ij})$$

where M_i represents the number of marginal records to annotate for class j , N is the total number of records to annotate.