IP-CRR: Information Pursuit for Interpretable Classification of Chest Radiology Reports

Yuyan Ge[1]*, Kwan Ho Ryan Chan[1], Pablo Messina[2], and René Vidal[1]

[1] University of Pennsylvania, Philadelphia, USA
[2] Pontifical Catholic University of Chile, CENIA, iHEALTH, Santiago, Chile

Abstract. The development of AI-based methods to analyze radiology reports could lead to significant advances in medical diagnosis, from improving diagnostic accuracy to enhancing efficiency and reducing workload. However, the lack of interpretability of AI-based methods could hinder their adoption in clinical settings. In this paper, we propose an interpretable-by-design framework for classifying chest radiology reports. First, we extract a set of representative facts from a large set of reports. Then, given a new report, we query whether a small subset of the representative facts is entailed by the report, and predict a diagnosis based on the selected subset of query-answer pairs. The explanation for a prediction is, by construction, the set of selected queries and answers. We use the Information Pursuit framework to select the most informative queries, a natural language inference model to determine if a fact is entailed by the report, and a classifier to predict the disease. Experiments on the MIMIC-CXR dataset demonstrate the effectiveness of the proposed method, highlighting its potential to enhance trust and usability in medical AI. Code is available at: `https://github.com/Glourier/MICCAI2025-IP-CRR`.

Keywords: Interpretable ML · Text classification · Radiology reports.

1 Introduction

When a doctor makes a decision that can potentially have life-changing consequences for a patient, the transparency of the decision-making is invaluable. As machine learning (ML) methods become more integrated into healthcare, there is a pressing need for safe and trustworthy models. For a prediction task, interpretability serves as a crucial aspect of model transparency that answers "why" and "how" a model made a prediction, fostering user trust.

Traditionally, interpretability is carried out by first training a black-box model that optimizes performance, and then producing an explanation in the form of heatmaps [30,6], selection of features [28,21], or more generally, attributions [33,31]. These post-hoc explanation methods are widely used in many areas such as computer vision, finance, and healthcare. However, these methods lack faithfulness [29], because the explanation could potentially be an artifact

---

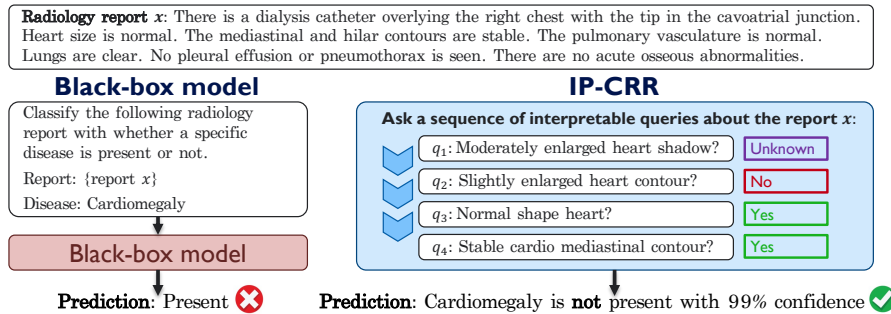* Corresponding author. Email: yyge@seas.upenn.edu

Radiology report x: There is a dialysis catheter overlying the right chest with the tip in the cavoatrial junction. Heart size is normal. The mediastinal and hilar contours are stable. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.

Black-box model

Classify the following radiology report with whether a specific disease is present or not.

Report: {report x}

Disease: Cardiomegaly

Black-box model

Prediction: Present ✗

IP-CRR

Ask a sequence of interpretable queries about the report x:

q_1: Moderately enlarged heart shadow?   Unknown

q_2: Slightly enlarged heart contour?   No

q_3: Normal shape heart?   Yes

q_4: Stable cardio mediastinal contour?   Yes

Prediction: Cardiomegaly is **not** present with 99% confidence ✓

Fig. 1: Illustration of the interpretable-by-design method.

of the explanation method rather than evidence of the model's underlying reasoning. To address this, interpretable-by-design models were proposed to make predictions solely based on a set of interpretable factors [16], which usually take the form of a set of concepts [18] or queries [9] about the input. For example, disease diagnosis for a patient can rely on a list of symptoms, which naturally explain the final prediction. While these methods improve faithfulness, they often require making careful and meticulous designs about the model architecture, optimization, and the set of interpretable concepts or questions. This complexity poses a significant challenge, particularly in the medical domain, where the need for both accuracy and interpretability are critical.

Take the task of Chest Radiology Report (CRR) classification as an example (Figure 1). A report typically contains a few key insights about the patient, and the task is to predict the associated diseases. For predictions to be interpretable, they must be grounded on a set of queries about the report or patient, such as "*Is the heart contour enlarged?*" or "*Does the lung volume decrease?*". This requires both the queries and their corresponding answers. Importantly, some queries may lack readily available answers in the report. This setting drastically differs from most settings in interpretable-by-design methods, where all answers are typically available and easily obtained using out-of-the-shelf models [8].

In this work, we propose an interpretable-by-design framework for CRR classification, which extends the Information Pursuit (IP) framework from image domain [9,7] to text. Our proposed IP-CRR method involves: 1) curating a set of interpretable queries by mining them from existing CRRs; 2) applying natural language inference (NLI) to determine if a fact is entailed by the report; and 3) learning a classifier that makes predictions by sequentially selecting a list of queries in order of information gain, answering those queries, and predicting the class from the sequence of query-answer pairs. Implementing our approach requires two key innovations relative to [9,7]: 1) modifying the IP framework to accommodate large language models (LLMs) for answering queries, and 2) handling "Unknown" answers, as illustrated in Figure 1. Moreover, our method achieves higher F1 score against black-box baselines such as Flan-T5-large and interpretable-by-design methods such as Concept Bottleneck Models. Qualita-

tively, the explanations provided by IP-CRR offer a potential venue for trustworthy and transparent predictions in clinical decision-making.

2 Related Work

Radiology Report Classification. Automatic disease diagnosis from radiology reports is a critical task that enables large-scale use of ML models in clinical settings. Natural language processing methods are widely used to extract diagnostic labels from free-text radiology reports. There are two main approaches for this task: rule-based labeling and ML-based labeling. Rule-based methods [5,25,4,14] rely on predefined rules or regular expressions to extract diseases from the report. However, these methods struggle with the complexity and variability of clinical language. ML-based methods [34,32,26,13] learn the patterns of text automatically without relying on handcrafted rules. For example, CheXbert [32] is a BERT-based medical domain specialized model pretrained on rule-based annotations, and finetuned on a small expert annotated dataset. However, despite their performance, they lack interpretability, making it hard for radiologists to trust their predictions in clinical use.

Interpretable-by-design Methods. Interpretable-by-design methods aim to produce explanations that are faithful to the model. State-of-the-art methods ensure this by using an intermediate layer of interpretable factors (e.g., concepts or queries) based on which a prediction is made. For example, Concept Bottleneck Models (CBMs) first map a given input to a list of concept scores, each corresponding to the presence of a concept in the input, and then *linearly* map concept scores to a predicted label. CBMs have found many successful applications in domains such as image classification and clinical prediction [23,10,35,17,24]. However, CBMs are sensitive to having concepts of very high quality and their use of a linear classifier limits their accuracy. In this work, we focus on another framework known as Information Pursuit (IP) [9,7,12], which provides interpretability via sequentially selecting queries in order of information gain, and then making predictions based on the sequence of query-answer pairs. Therefore, IP aligns more closely with the interactive question-answering process between a doctor and a patient. The challenge, however, involves constructing a set of meaningful queries and obtaining their answers. Currently, it has only been explored for image classification, where [8] leverages LLMs [1] and vision-language models [19]. In this work, we aim to leverage natural language inference for text and medical domain. We discuss the framework in detail in Section 3.3.

3 Method

In this section, we describe IP-CRR, the proposed interpretable-by-design model for classifying CRRs. Our pipeline consists of three main components: 1) generate a set of queries from a large number of reports (Sec. 3.2), 2) design a mechanism to answer queries automatically (Sec. 3.2), and 3) integrate the Information Pursuit framework (Sec. 3.3) to conduct interpretable classification.
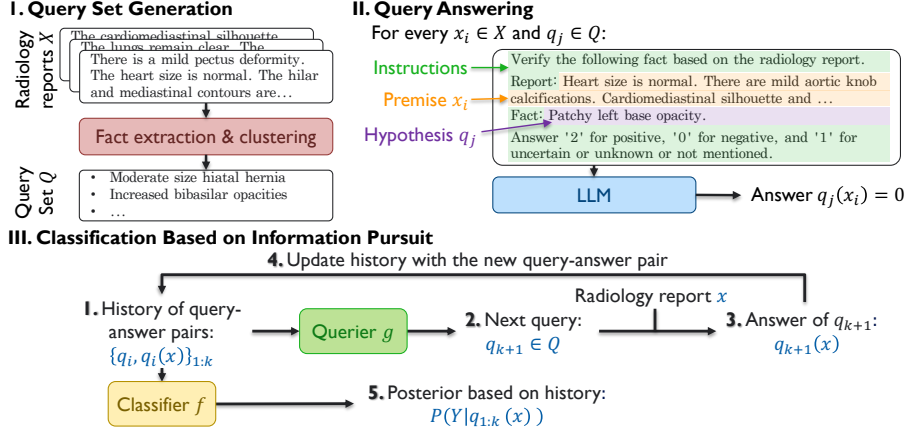
Fig. 2: Framework of IP-CRR, which consists of three parts: 1) query set generation by extracting and clustering facts from a large-scale radiology report dataset, 2) query answering based on natural language inference, and 3) interpretable classification based on variational information pursuit.

3.1   Problem Formulation

We consider the problem of classifying a CRR based on a sequence of interpretable queries. Let $X \in \mathcal{X}$ be a random variable representing a CRR, where $\mathcal{X}$ is the space of all possible CRRs, and $Y \in \mathcal{Y} = \{0, 1\}$ denotes the corresponding binary label. Our goal is to determine Y given X by sequentially querying X with a set Q of predefined, interpretable queries.

Each query $q \in Q$ is a function $q : \mathcal{X} \to \mathcal{A}$, where $\mathcal{A} = \{-1, 0, 1\}$ represents the possible answers: negative, unknown, and positive. For example, given a radiology report x and a query $q =$ "*Is there a decreased opacity on the right?*", $q(x) = 1$ implies there is a decreased opacity on the right, $q(x) = -1$ implies there is not, and $q(x) = 0$ implies the query cannot be answered from the given report. We assume the query set Q is interpretable, task-specific, and sufficient[1].

The classification process involves selecting the most informative queries iteratively until sufficient information is obtained to predict Y with high confidence.

3.2   Constructing and Answering Queries from Existing Databases

Our final goal is to classify whether a certain disease is present or not in a given CRR. To apply the IP framework, we first need to define a set of interpretable and task-relevant queries about the report. Each query corresponds to a question about whether a specific fact is "true", "false", or "unknown" given the report. For example, the fact "*normal heart shape*" can be thought of as the query "*Is the heart shape normal?*" While there are many choices for the query set, we argue

---

[1] That is, $P(Y \mid x) = P(Y \mid \{x' \in \mathcal{X} : q(x) = q(x'), q \in Q\})$. See [9].

that curating a sufficiently large set of facts as queries for CRR classification allows us to build interpretable models with good predictive performance. Next, we describe how to construct this query set and answer queries for a CRR.

Queries q**.** The procedure for obtaining a large set of representative, domain-specific, and interpretable queries is illustrated in Figure 2 Part I. We construct the query set Q by mining facts from a large-scale CRR dataset using a recently developed fact extraction model [22]. Specifically, we analyze over 220k CRRs from the MIMIC-CXR dataset [15] and parse them into more than 660k sentences using the Natural Language Toolkit [2]. Next, we extract over 590k facts from these 660k sentences by first prompting the GPT-3.5 and GPT-4 models to generate facts for a subset of the sentences and then using these sentence-fact pairs as training data to finetune the T5-small [27] sequence-to-sequence model for fact extraction. After fact extraction, we encode each fact into a 128-dimension latent space using the fact encoder CXRFE [22]. We apply k-means clustering to all 590k extracted facts in the fact embedding space, then select the nearest fact to each cluster center, and conduct deduplication if any two selected facts have a cosine similarity greater than 0.97. After this process, we obtain 520 representative facts, where each of them is a short fact describing the clinical findings, e.g., "*moderate size hiatal hernia*" or "*increased bibasilar opacities*".

Query Answers $q(x)$**.** Ideally, query answers should be obtained from the report through annotations made by radiologists. However, manual annotation is infeasible due to the large number of reports and queries. Moreover, in contrast to image classification, it often occurs that a query cannot be answered based on the provided radiology report. For example, if the description of the heart is not mentioned in the report, we cannot answer the query "Is the heart size normal?" from the report, although the heart is shown in the image. To obtain the query answers in a scalable and accurate manner, we propose to leverage natural language inference (NLI). Specifically, we pass both the query and the CRR as inputs to the Flan-T5-large [11] LLM and ask it to predict the query answer as either "yes", "no" or "unknown". An example of the prompt is illustrated in Figure 2 Part II. The prompt consists of three parts: instruction, premise, and hypothesis. The instruction serves as guidance to the model and is fixed for every sample, the premise is the radiology report, and the hypothesis is the fact to be verified as true for the given report. For every report and query, the original model output is 0, 1, or 2 for negative, unknown, or positive. Then we normalize them to form the query answer $q(x) \in \{-1, 0, 1\}$.

3.3   Classification Based on Variational Information Pursuit

IP. We aim to classify a report using Information Pursuit (IP), an interpretable-by-design framework that sequentially selects queries about the data in order of information gain. Let $I(q(X); Y)$ denote the mutual information between $q(X)$ and Y. Given a sample CRR $x^{\mathrm{obs}} \in \mathcal{X}$, IP selects the queries as follows:

$$q_1 = \underset{q \in Q}{\operatorname{argmax}} I(q(X); Y); \quad q_{k+1} = \underset{q \in Q}{\operatorname{argmax}} I(q(X); Y \mid q_{1:k}(x^{\mathrm{obs}})). \qquad (1)$$

Here, $q_{k+1} \in Q$ refers to the new query selected at step $k+1$ based on the history of query-answer pairs $q_{1:k}(x^{\mathrm{obs}})$. The algorithm terminates at iteration L when there are no more informative queries, i.e., $\forall q \in Q, I(q(X); Y \mid q_{1:L}(x^{\mathrm{obs}})) = 0$. Furthermore at each iteration k, one can also compute the posterior distribution (i.e. the prediction) $P(Y \mid q_{1:k}(x^{\mathrm{obs}}))$. To carry out IP, [9] proposed to learn a VAE to model the joint distribution of all query answers and labels $P(Q(X), Y)$. However, learning a good generative model is challenging when the size of Q is large, so a variational approach is proposed as an alternative.

V-IP. Rather than modeling the joint distribution of queries and labels to directly compute the most informative queries, Variational Information Pursuit (V-IP) [7,8,12] finds a querier g that selects the most informative query q given any random history S of query-answer pairs, and a predictor f that models the posterior distribution of Y given q and S, by optimizing the following objective:

$$
\begin{aligned}
\min_{f,g} \quad & \mathbb{E}_{X,S}\left[D_{KL}(P(Y \mid X) \parallel \hat{P}(Y \mid q(X), S)) \right] \\
\text{where} \quad & q := g(S) \in Q \\
& \hat{P}(Y \mid q(X), S) := f(\{q, q(X)\} \cup S).
\end{aligned}
\tag{V-IP}
$$

The variational objective aims to minimize the KL-divergence between the true posterior distribution and the posterior distribution given any history and an additional query and its answer. In practice, the classifier and querier are parameterized as neural networks, f_θ and g_η, making them efficient for carrying out the IP algorithm and scalable for evaluating large-scale datasets. The history S is represented using a masked vector with length $|Q|$. Since our task is classification, where Y is a categorical variable, the KL-divergence is empirically equivalent to the cross-entropy loss. Once f_θ and g_η are trained, the querier g_η serves as a function that directly outputs the query with the highest mutual information [7], hence IP can be directly carried out by having $q_{k+1} = g(\{q_{1:k}, q_{1:k}(x)\})$.

To implement IP-CRR, we first obtain a query set from large amounts of CRRs as described in Sec. 3.2, then answer each query for each report with an LLM. Finally, we apply V-IP to train a querier and a classifier to generate a sequence of interpretable query-answer pairs for the final classification.

4   Experiments

In this section, we will demonstrate the effectiveness of IP-CRR at performing CRR classification.

4.1   Dataset and Implementation Details

Dataset. We evaluate the proposed method on a public radiology dataset named MIMIC-CXR [15]. In this dataset, each patient is called "subject", and each examination is called "study". Each study has one radiology report and multi-label ground-truth annotations [13] indicating whether diseases are present. The dataset is randomly split into training, validation, and test sets in a 7:1:2 ratio.

Table 1: Predictive performance of IP-CRR and baseline models across six tasks, measured by Average precision (AP, left) and F1 score (right). Higher values indicate better performance. The best results among black-box models are <u>underlined</u>, while the best interpretable model results are **bolded**.

Methods	AP						F1					
	LO	CA	SD	CM	PE	PN	LO	CA	SD	CM	PE	PN
CXR-BERT (FT-Last)	0.900	0.361	0.969	0.864	0.945	0.449	0.829	0.223	0.912	0.789	0.887	0.449
CXR-BERT (FT-All)	<u>0.984</u>	<u>0.992</u>	<u>0.970</u>	<u>0.964</u>	<u>0.962</u>	<u>0.641</u>	<u>0.987</u>	<u>0.991</u>	<u>0.978</u>	<u>0.982</u>	<u>0.953</u>	<u>0.541</u>
Flan-T5-large	0.527	0.073	0.445	0.380	0.616	0.190	0.663	0.139	0.321	0.543	0.754	0.299
CBM	0.947	0.345	0.934	0.791	0.874	0.432	0.884	0.241	0.853	0.738	0.801	0.431
IP-CRR	**0.972**	**0.578**	**0.959**	**0.892**	**0.925**	**0.468**	**0.918**	**0.350**	**0.889**	**0.811**	**0.860**	**0.451**
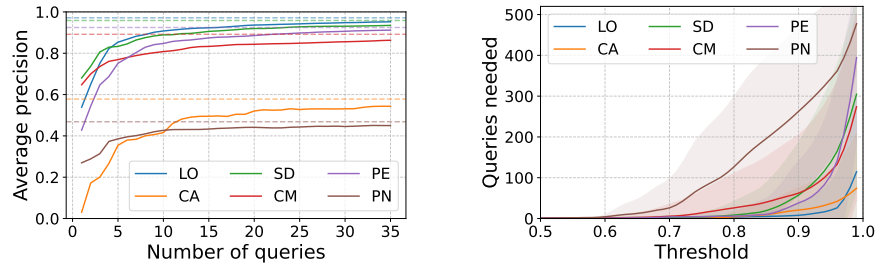
(a) Average precision versus number of queries asked. Dashed lines mark the maximum performance over all queries.

(b) Number of queries needed over thresholds on the posterior. Shaded regions indicate the variance across runs.

Fig. 3: Quantitative evaluation of IP-CRR as the number of queries increases.

Implementation Details. The architecture of both the querier and the classifier consists of a multilayer perceptron with 5 layers (see code for details). The IP-CRR model terminates with a confidence threshold of 0.85 or a maximum number of queries of 200. The optimizer is AdamW [20] with an initial learning rate of 0.0001 and a cosine annealing decay. The model is trained with a batch size of 128 for 1000 epochs. The loss function is weighted binary cross-entropy to alleviate the class imbalance. All experiments are implemented in PyTorch 2.5.1 using an NVIDIA A5000 GPU with a memory of 24.5GB.

4.2 Quantitative Results and Comparisons

We evaluate the performance of IP-CRR with all queries on six binary classification tasks: Lung Opacity (LO), Calcification of the Aorta (CA), Support Devices (SD), Cardiomegaly (CM), Pleural Effusion (PE), and Pneumonia (PN). Baselines include CXR-BERT [3] finetuned on the last layer (FT-Last) and on all layers (FT-All), Flan-T5-large [11] without fine-tuning, and CBM [18]. Notably, the first three are black-box models and CBM is interpretable by design.

As shown in Table 1, IP-CRR outperforms CXR-BERT (FT-Last), Flan-T5-large, and CBM in average precision (AP) and F1 score on the LO, CA, CM, and PN tasks, and achieves comparable performance on SD. However, compared to CXR-BERT (FT-All), IP-CRR matches its AP on LO, but underperforms on other tasks. We attribute this gap primarily to the CXR-BERT's black-box nature, its more complex BERT-based architecture, and the fact that all its parameters are fine-tuned during training. In contrast, IP-CRR is implemented using a simple MLP with significantly fewer parameters, prioritizing interpretability.

Furthermore, in Figure 3a, we evaluate AP as a function of the number of queries. For the LO task, IP-CRR achieves 0.95 AP with just 30 queries, and 0.97 with 100 – fewer than 20% of the total of 520 queries. Similar trends hold across other tasks, indicating that IP-CRR generally selects informative queries in its early iterations. Alternatively, we observe that IP-CRR converges to the maximum performance with only a small subset of queries. Compared to the baselines in Table 1, IP-CRR achieves better performance while using a smaller number of queries than CBMs, which use all available queries.

Next, in Figure 3b, we plot the number of queries needed as the threshold on the posterior changes. The threshold on the posterior approximately measures the confidence the user requires the predictor to reach before terminating. We observe that the IP-CRR predictor requires fewer than 150 queries to achieve a confidence of 0.95 for most of tasks. This demonstrates that IP-CRR is able to find informative queries to make confident predictions.

4.3   Qualitative Analysis of Query-answer Chains

The core mechanics of using IP-CRR for interpretable predictions is the iterative process of selecting informative queries. This allows the user to observe how the posterior evolves as the number of queries increases. In Figure 4, we demonstrate two examples of query-answer chains obtained using IP-CRR. For the example on the left, our method first selects $q_1 = $ "*Opacification likely reflects atelectasis?*". From the report, the model infers that $q_1(x)$ is not present in the report. Then, it selects the query $q_2 = $ "*Patchy opacity at right lung base?*". Continuing in this fashion, after 9 queries, the model concludes that the patient has LO. The same sequential deduction can be observed in the example on the right, highlighting the interpretability IP-CRR offers when applied to CRRs.

5   Conclusion

In this work, we have proposed an interpretable-by-design framework for CRR classification. We obtain a large set of queries by mining them from existing CRRs, and answer them using NLI. To evaluate the proposed framework, we conducted experiments on a large-scale CRR dataset and on multiple tasks, with both quantitative and qualitative results that illustrate interpretability and competitive performance. A current limitation is the reliance on a general-domain language model for query answering. We leave the use of stronger medical foundation models for generating more expressive query answers to future work.
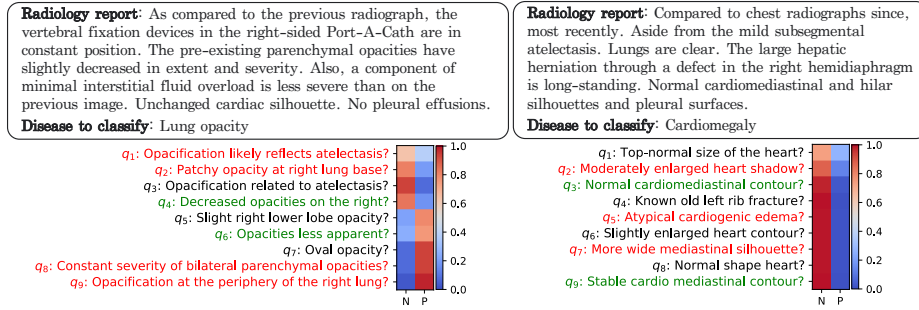
Fig. 4: Examples of query-answer chains from IP-CRR. Each row in the colored matrix represents the posterior $P(Y \mid q_{1:k}(x))$ after selecting the query q_k. Text in red means the answers is "no", green means "yes", and black means "unknown". "N" or "P" means negative or positive prediction of the disease, respectively.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., et al.: GPT-4 technical report. In: arXiv:2303.08774 (2024)
2. Bird, S., Loper, E.: NLTK: The natural language toolkit. In: ACL (2004)
3. Boecking, B., Usuyama, N., Bannur, S., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV (2022)
4. Bozkurt, S., Alkim, E., Banerjee, I., Rubin, D.L.: Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. J. Digit. Imaging (2019)
5. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. (2001)
6. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)
7. Chattopadhyay, A., Chan, K.H.R., Haeffele, B.D., Geman, D., Vidal, R.: Variational information pursuit for interpretable predictions. In: ICLR (2023)
8. Chattopadhyay, A., Chan, K.H.R., Vidal, R.: Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In: ICLR (2024)

9. Chattopadhyay, A., Slocum, S., Haeffele, B.D., Vidal, R., et al.: Interpretable by design: Learning predictors by composing interpretable queries. IEEE TPAMI (2022)
10. Chowdhury, T.F., Phan, V.M.H., Liao, K., et al.: AdaCBM: An adaptive concept bottleneck model for explainable and accurate diagnosis. In: MICCAI (2024)
11. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., et al.: Scaling instruction-finetuned language models. JMLR (2024)
12. Covert, I.C., Qiu, W., Lu, M., Kim, N.Y., White, N.J., Lee, S.I.: Learning to maximize mutual information for dynamic feature selection. In: ICML (2023)
13. Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.H., et al.: Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. Med. Image Anal. (2024)
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
15. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data (2019)
16. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: ICML (2018)
17. Kim, E., Jung, D., Park, S., Kim, S., Yoon, S.: Probabilistic concept bottleneck models. arXiv:2306.01574 (2023)
18. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML (2020)
19. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
20. Loshchilov, I.: Decoupled weight decay regularization. arXiv:1711.05101 (2017)
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. NeurIPS (2017)
22. Messina, P., Vidal, R., Parra, D., Soto, A., Araujo, V.: Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. In: ACL Findings (2024)
23. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: ICLR (2023)
24. Panousis, K.P., Ienco, D., Marcos, D.: Sparse linear concept discovery models. In: ICCV (2023)
25. Peng, Y., Wang, X., Lu, L., et al.: NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Transl. Sci. Proc. (2018)
26. Pereira, S.C., Mendonça, A.M., Campilho, A., Sousa, P., Lopes, C.T.: Automated image label extraction from radiology reports—a review. Artif. Intell. Med. (2024)
27. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" explaining the predictions of any classifier. In: KDD (2016)
29. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. (2019)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
31. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML (2017)

32. Smit, A., Jain, S., Rajpurkar, P., et al.: CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: EMNLP (2020)
33. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
34. Wang, S., Lin, M., Ding, Y., Shih, G., Lu, Z., Peng, Y.: Radiology text analysis system (RadText): architecture and evaluation. In: ICHI (2022)
35. Yang, Y., Panagopoulou, A., Zhou, S., et al.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: CVPR (2023)