TriAID: Chest X-Ray Follow-Up Imaging Recommendation

Edward Tanurkov
Queen's University
edward.tanurkov@queensu.ca

Theo Leone
Queen's University
23ltf@queensu.ca

Shravan Agnihotri
Queen's University
22pj42@queensu.ca

Leila Salem
Queen's University
leila.salem@queensu.ca

Abdullah Mohsin
Queen's University
abdullah.m@queensu.ca

Alex Kent*
Queen's University
22bmb7@queensu.ca

Travis (Jeaneon) Hong*
Queen's University
23dn20@queensu.ca

***Abstract*—This paper presents TriAID, a clinical decision-support framework for identifying follow-up imaging recommendations from chest radiography studies. The proposed system combines structured report-derived features with pathology probability outputs generated by the MedRAX computer vision model to predict follow-up imaging categories. A One-vs-Rest (OvR) multi-label classification approach is employed to classify radiology report information into recommendation categories including X-ray, CT, or no follow-up. The model is trained on 26,797 chest radiography studies paired with corresponding clinical reports extracted from the publicly available ReXGradient-160K dataset. A pre-processing pipeline was developed to identify follow-up recommendations from radiologist findings and impressions, enabling construction of a structured training dataset. Multiple feature extraction and classification strategies were evaluated to determine an effective modelling approach. The resulting model achieves a micro-averaged precision of 78.56%, recall of 82.29%, and F1-score of 80.38%. The proposed framework is intended to assist clinical workflows by providing structured follow-up recommendation predictions while maintaining interpretability. The system is designed as a supportive tool for clinical environments and is not intended to replace professional medical judgment.**

I. Introduction

Chest radiography remains one of the most frequently performed diagnostic imaging procedures and plays a central role in the evaluation of cardiopulmonary disease. Radiology reports generated from these studies often include recommendations for follow-up imaging to monitor disease progression, clarify uncertain findings, or evaluate treatment response. Accurate identification of these recommendations is important for ensuring appropriate patient management and continuity of care.

Recent advances in machine learning have demonstrated strong performance in medical image analysis and radiology report interpretation. However, much of the existing work has focused primarily on diagnostic classification or automated report generation. Comparatively less attention has been directed toward identifying and structuring follow-up imaging recommendations within radiology documentation.

This work presents TriAID, a multimodal clinical decision-support framework designed to identify follow-up imaging recommendations associated with chest radiography studies. The proposed system integrates structured report-derived features with pathology probability estimates produced by a computer vision model, enabling prediction of follow-up imaging categories through a multi-label classification approach. In addition to the predictive model, the system is integrated into an application framework that enables structured access to recommendations and supporting clinical context.

By framing follow-up imaging recommendation as a structured multi-label prediction task, the proposed approach aims to support consistency in recommendation identification while maintaining interpretability and practical applicability within clinical workflows.

A. Motivation

Although chest radiography is widely used in clinical practice, follow-up imaging recommendations contained within radiology reports may vary across institutions and practitioners. This variability is particularly evident in cases involving subtle findings or uncertain diagnoses, such as suspected pneumonia or indeterminate thoracic abnormalities [1], [2]. Differences in interpretation, reporting style, and institutional practices may contribute to inconsistency in follow-up guidance.

Prior studies have examined the outcomes and diagnostic value of follow-up imaging recommendations derived from radiographic findings [2], [3]. These studies highlight both the importance of follow-up imaging in clinical decision-making and the challenges associated with determining appropriate follow-up strategies.

In practice, follow-up decisions frequently depend on multiple factors, including imaging interpretation, patient demographics, clinical indications, and physician judgment. As a result, identifying and structuring follow-up recommendations within radiology reports can be complex.

While artificial intelligence methods have been widely applied to diagnostic imaging tasks, comparatively limited work has focused on extracting and standardizing follow-up imaging recommendations from radiology documentation. Developing a structured and data-driven approach for identifying these recommendations may support improved consistency in clinical

workflows and assist practitioners in reviewing radiology report information.

B. Related Works

Machine learning has been widely applied within radiology research, particularly in areas such as abnormality detection, image classification, and automated report generation. In the context of chest radiography, numerous studies have demonstrated strong performance in identifying pathological findings using deep learning models trained on large-scale medical imaging datasets [4].

More recently, multimodal frameworks have been developed to integrate imaging data with structured clinical information in order to support more comprehensive medical reasoning. One such example is the MedRAX framework, which combines radiographic features with structured metadata to enhance clinical interpretation capabilities [5]. These approaches demonstrate the potential of multimodal models for improving radiology-related prediction tasks.

Despite these advances, most existing systems emphasize diagnostic classification or report interpretation rather than the identification of follow-up imaging recommendations. Radiology literature has documented variability in follow-up imaging practices across institutions and patient populations [3], [6]. However, relatively limited research has formalized follow-up recommendation identification as a machine learning prediction task.

The present work addresses this gap by explicitly modeling follow-up imaging recommendation as a structured multi-label classification problem. By integrating structured clinical features with image-derived pathology probabilities, the proposed system aims to support identification of follow-up imaging recommendations within radiology reports while maintaining compatibility with practical clinical workflows.

C. Problem Definition

We formulate follow-up imaging recommendation as a structured multi-label classification problem. Given structured patient metadata derived from chest radiography encounters, including demographic attributes and clinical notes, the objective is to predict one or more clinically appropriate follow-up imaging categories.

Let $X \in \mathbb{R}^d$ denote the structured feature representation extracted from patient data, and let $y \in \{0,1\}^k$ represent the vector of follow-up recommendation labels across k possible categories. The goal is to learn a mapping function $f : X \to y$ that produces consistent, interpretable, and clinically meaningful recommendations.

Building upon the MedRAX framework [5], [7], we integrate domain-informed pre-processing with an OvR classification strategy to generate multi-label predictions. Unlike prior radiographic reasoning systems that emphasize diagnosis or report generation, our formulation targets standardized follow-up recommendation aligned with clinical workflow constraints. The model produces multi-label outputs accompanied by calibrated probability estimates and confidence scores, enabling transparent assessment of recommendation strength.

Additionally, LLM functionality is incorporated to provide contextual explanatory reasoning alongside quantitative predictions. This structured yet interpretable framework supports reproducibility, systematic evaluation, and practical integration into clinical decision-support systems.

II. Methodology

This section describes the dataset, preprocessing procedures, model development process, and evaluation methodology used to construct the proposed follow-up recommendation system.

A. Data

Data Source

To ensure compliance with ethical and privacy considerations within the medical domain, publicly available de-identified datasets were explored. The dataset selected for this study was *ReXGradient-160K*, which contains approximately 160,000 chest radiography studies paired with corresponding radiological reports [8]. These reports originate from 109,487 unique de-identified patients across three U.S. health systems and 79 medical sites, collected between March 2004 and September 2024.

The radiological reports are provided in structured CSV format, where each report is pre-parsed from original radiology documentation and associated with its corresponding imaging study. Each study contains at least one radiographic image, with multiple images representing distinct anatomical views rather than repeated acquisitions.

Selected Features

From the structured report data, several attributes were identified as particularly relevant for follow-up recommendation modeling. The following features were selected as primary inputs for model training:

- Patient age (ranging from 0 to 90 years, with a mean of 45.77)
- Patient sex (F, O, M)
- Clinical indication, consisting of a short textual description summarizing the patient's presenting symptoms

Although the dataset also contained *Findings* and *Impression* fields, these were excluded from the training features after preliminary experimentation indicated that their inclusion degraded model performance. Additional dataset attributes were omitted in order to focus on clinically interpretable features. The construction of the target variable is described in the following subsection.

B. Model Training and Architecture

Pre-processing Techniques

The original *ReXGradient-160K* dataset did not explicitly contain a column indicating recommended follow-up imaging. However, follow-up recommendations are frequently described within the *Findings* and *Impression* sections of radiology

reports. Consequently, a preprocessing pipeline was developed to extract these recommendations from the textual reports.

A Python-based parsing script was implemented to identify common vocabulary patterns, sentence structures, and semantic indicators associated with follow-up imaging recommendations. When a recommendation was detected, the relevant sentence was stored in a *FollowUpEvidence* column, while the extracted recommendation category was recorded in a *FollowUpTests* column. The latter served as the target variable for model training.

Following this extraction process, the effective dataset size was reduced from 160,000 studies to 26,849 observations containing identifiable follow-up recommendations.

The extracted recommendations included follow-up imaging modalities such as X-ray, CT, ultrasound, and MRI, as well as cases where no follow-up imaging was recommended. In some instances, multiple imaging modalities were recommended simultaneously, although no cases were observed where more than two modalities were recommended together.

Subsequent preprocessing steps were applied to the selected features. Patient age values were converted from string-based representations (e.g., "45Y" for 45 years or "32W" for 32 weeks) into numerical values expressed in years. The categorical sex variable was encoded using one-hot encoding. Clinical indication text was transformed into numerical features using Term Frequency–Inverse Document Frequency (TF–IDF) vectorization with both unigram and bigram representations, enabling capture of contextual patterns within the free-text descriptions [9].

During preprocessing, a single observation with missing age information was removed, resulting in 26,848 observations.

The distribution of follow-up recommendation classes exhibited substantial imbalance. After filtering and preprocessing, four primary recommendation categories were retained:

TABLE I
FINAL FEATURE COLUMN CLASS SPLIT

Classes	No Follow-Up	X-Ray	CT	X-Ray + CT
Values in Class	22,001	2731	1893	172

The original dataset contained nine distinct recommendation classes, including ultrasound and MRI modalities. However, these classes contained extremely small numbers of observations (24 ultrasound, 20 MRI, and seven additional combined categories). Due to insufficient sample size for effective model training, these classes were excluded from the dataset.

Following this filtering process, the final dataset consisted of 26,797 observations.

A train–validation–test split of 60/20/20 was employed. This partitioning was selected to balance training data availability with reliable model evaluation while mitigating the risk of overfitting, consistent with common dataset partitioning practices for moderately sized datasets [10]. Class imbalance was intentionally preserved to reflect the underlying distribution of real-world radiology recommendations.

In addition to the structured report features, the associated radiographic images were processed using the MedRAX framework to generate probabilities for 18 pathology classes [5]. These probabilities were incorporated as additional feature columns within the model input representation.

Baseline Models Tested

Two baseline models were evaluated to assess suitable classification architectures for the task.

The first model employed an OvR strategy implemented using the scikit-learn library. OvR classification constructs an independent binary classifier for each label, distinguishing that label from all other classes [11]. This approach provides computational efficiency while maintaining a high level of interpretability, which is particularly desirable in medical decision-support systems.

In the present setting, three independent classifiers were trained corresponding to the labels *X-ray*, *CT*, and *No Follow-Up*. The combined recommendation class *X-ray + CT* was modeled implicitly through multi-label prediction, allowing both relevant classifiers to activate simultaneously.

The second baseline model evaluated was a Random Forest classifier implemented using scikit-learn [12]. Random Forest models employ ensemble averaging across multiple decision trees and are often effective for heterogeneous feature spaces. However, their interpretability is typically lower than that of linear classification models.

The micro-average results of the baseline models are presented in Table II.

TABLE II
BASELINE MODEL MICRO AVERAGE METRIC COMPARISON

Model	Precision	Recall	F1
One-vs-Rest	98%	99%	98%
Random Forest	87%	80%	83%

These experiments used baseline configurations without hyperparameter optimization or class-specific probability thresholds. Given the substantially stronger performance and greater interpretability of the OvR approach, this architecture was selected for subsequent development.

An additional CatBoost-based model was briefly explored to evaluate potential improvements from gradient-boosted decision trees. However, due to project time constraints, this model was not fully evaluated and therefore is not included in the final analysis.

Chosen Model and Configurations

Following baseline evaluation, the OvR classifier was further refined through threshold tuning and configuration adjustments. Model optimization focused primarily on maximizing the F1-score, which balances precision and recall.

Because the classifier produces independent probability estimates for each label, class-specific probability thresholds were introduced to determine whether a recommendation should be issued. Three primary recommendation buckets were defined: *X-ray*, *CT*, and *No Follow-Up*. As previously noted,

simultaneous recommendations for X-ray and CT arise naturally through multi-label prediction rather than through a separate classification category.

The final probability thresholds selected for each recommendation bucket were:

- X-ray: 20%
- CT: 20%
- No Follow-Up: 40%

These thresholds were determined through empirical experimentation across multiple configurations of the OvR model. Initial optimization focused on maximizing recall, as false negatives—cases where follow-up imaging is recommended but not identified by the model—represent a particularly undesirable outcome in clinical settings. However, balancing recall with precision ultimately produced more reliable and interpretable recommendations.

Architecture

The overall system architecture consists of three primary layers: a Training Layer, a Model Layer, and an Application Layer, as illustrated in Figure 1.

The *Training Layer* is responsible for model development and retraining when necessary. This layer incorporates the data sources described previously and performs preprocessing, feature extraction, and model optimization.

The *Model Layer* contains the trained OvR classifier and associated feature processing components. Structured clinical report features are combined with pathology probability outputs generated by the MedRAX framework from chest radiography images. These combined features are used as inputs to the classification model to generate follow-up imaging recommendations.

Predicted recommendations are subsequently transmitted to the *Application Layer*, where they can be viewed through a user interface and optionally stored within a database for future retrieval. Storing previously computed recommendations allows the system to avoid recomputation when the same radiography study is accessed again.

The Application Layer also includes a conversational assistant designed to provide contextual explanations for model outputs. This component utilizes a retrieval-augmented generation (RAG) approach in which text embeddings generated using the HuggingFace `sentence_transformers all-MiniLM-L6-v2` model are used to retrieve relevant clinical guidance from American College of Radiology (ACR) resources [13], [14]. Retrieved context is then provided to a language model via the DeepSeek API to generate explanatory responses [15].

This assistant component is designed to provide supplementary contextual information and is not intended to replace clinical expertise. The primary focus of evaluation in this work remains the predictive performance of the follow-up recommendation model.

C. Evaluation Method

Model performance was evaluated primarily using micro average metrics focusing on recall and F1-score. These metrics were selected due to their suitability for imbalanced classification problems and their relevance to clinical decision-support systems, where both detection capability and prediction reliability are important.

Micro averaging was utilized as it aggregates the contributions of all classes before computing the metric, more specifically, the total number of true positives, false positives, and false negatives across all classes are summed and the metric is computed once using these global totals [16]. This approach effectively gives equal weight to each individual observation, meaning that classes with more samples contribute more strongly to the final metric, thus making it the metric of choice when imbalanced datasets are necessary for training [16].

Model configurations were compared by adjusting feature inputs and probability thresholds within the OvR classification framework. Quantitative evaluation was conducted using the held-out validation and test datasets.

In addition to quantitative metrics, preliminary qualitative inspection of model predictions was performed during development. Selected samples were manually reviewed to verify whether predictions appeared clinically plausible based on the corresponding imaging findings and patient indications. While this exploratory process helped identify potential configuration issues during development, it was not used as a formal evaluation metric due to its subjective nature.

III. Results

The performance of the proposed follow-up recommendation model was evaluated using standard classification metrics, including precision, recall, and F1-score. Because the task is formulated as a multi-label classification problem, performance was assessed using micro-averaged, macro-averaged, and weighted-averaged metrics to provide a comprehensive evaluation of model behavior across classes.

Table III presents the per-class performance of the model after applying the selected decision thresholds.

TABLE III
PER-CLASS CLASSIFICATION PERFORMANCE

Class	Precision	Recall	F1-score	Support
X-ray	38.94%	56.97%	46.26%	581
CT	36.36%	45.41%	40.39%	414
None	91.48%	89.09%	90.27%	4409

Overall model performance is summarized using aggregated evaluation metrics. The model achieved a micro-averaged precision of 78.56%, recall of 82.29%, and F1-score of 80.38%. Macro-averaged performance values were lower, with a precision of 55.59%, recall of 63.82%, and F1-score of 58.97%. Weighted averages yielded a precision of 81.61%, recall of 82.20%, and F1-score of 81.72%.

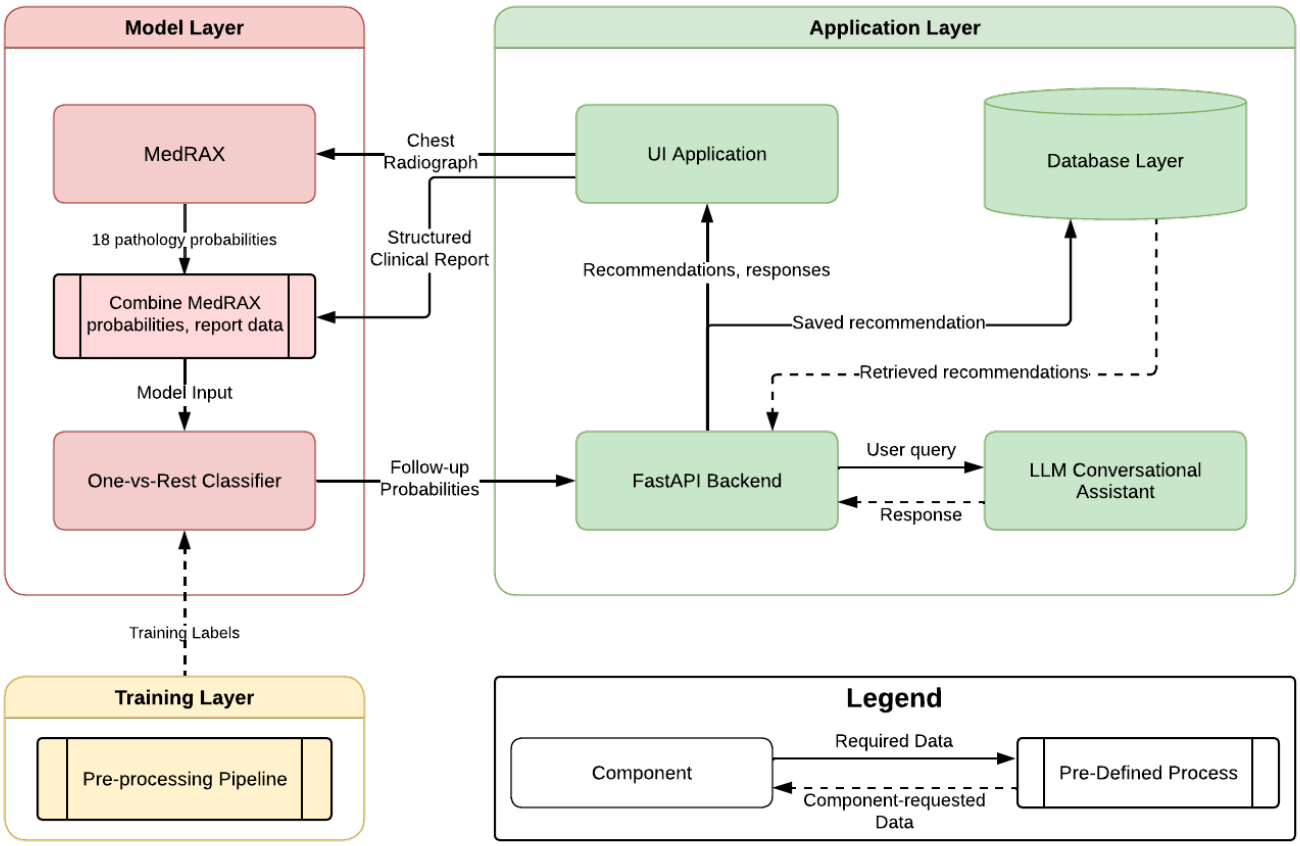

Fig. 1. TriAID Model and Application Architecture

The difference between micro-averaged and macro-averaged metrics reflects the class imbalance present in the dataset, where the majority of samples correspond to cases without follow-up imaging recommendations. Micro-averaging aggregates predictions across all samples and therefore reflects overall system performance, whereas macro-averaging assigns equal weight to each class regardless of frequency [16].

An additional factor to consider is the aforementioned thresholding, which affects the variance within the report values as it is an additional hyperparameter that required tuning. Each bucket was tuned to the final probability thresholds mentioned in the Chosen Model and Configurations subsection of the Methodology section.

Despite these challenges, the model demonstrates the ability to identify follow-up imaging recommendations with moderate accuracy while maintaining strong performance in distinguishing cases without follow-up recommendations. These results suggest that the proposed approach may provide useful structured signals for supporting clinical review of radiology reports.

The following figures is are example radiographs, with Table IV being the associated simplified clinical report inputs for each, along with the follow-up recommendations, confidence and reasoning evidence generated by the model reported in Table V.

Fig. 2. Sample radiograph 1

Fig. 3. Sample radiograph 2

TABLE IV
SAMPLE STRUCTURED CLINICAL REPORTS

Sample	Patient Age	Patient Sex	Indication
Radiograph 1	69 y.o.	F	respiratory failure
Radiograph 2	73 y.o.	F	short of breath

TABLE V
SAMPLE FOLLOW-UP RECOMMENDATIONS

Sample	Recommendation	Confidence
Radiograph 1	CT Chest	62%
Radiograph 2	Chest X-Ray (PA + Lateral)	80%
Evidence (MedRAX findings)		
Radiograph 1	Fibrosis, nodule, mass, pleural thickening, emphysema	
Radiograph 2	Hernia, lung opacity, mass, infiltration, fibrosis	

The above are two examples of successful recommendations by the model architecture, matching that of the associated labels. The confidence level is generated from the OvR classifier as its probability that the label is true, and the evidence is what the MedRAX classifier found as potential physical indicators based solely off the provided radiographs.

IV. DISCUSSION

A. Potential Inaccuracies in Indication Modelling

While self-developed pre-processing techniques, as mentioned in the Methodology section, enable structured representation of clinical indication, variability in free-text documentation introduces potential noise into the feature space. Differences in physician writing style, terminology specificity, and contextual omission may reduce semantic consistency across samples. Consequently, indication-derived embeddings may not uniformly reflect the true diagnostic intent.

Although domain-adapted text recognition programs improve contextual understanding, they remain sensitive to distributional shifts. This limitation may propagate uncertainty into downstream classification outputs. Future work may incorporate controlled vocabulary mapping or ontology-constrained preprocessing to mitigate semantic variability.

B. Notable Implementation Challenges

Several technical constraints were encountered during system development. Notably, MedRAX compatibility required implicit rescaling of input images to 8-bit resolution, as otherwise images not matching this resolution are silently setting all 18 pathology classes as 100% missing, thereby not attempting to classify anything and rejecting the image. This undocumented pre-processing requirement introduced an additional normalization step to ensure consistent inference behaviour.

Data accession and pre-processing consistency also required careful validation. Ensuring reproducible train–test separation from a single source dataset was critical to prevent data leakage and preserve structured evaluation integrity. These implementation details, while not algorithmically central, were essential for maintaining system reliability.

C. Additional System Components

Although the core contribution centers on multi-label follow-up recommendation modeling, the system is embedded within a deployable application, first mentioned in the Methodology section.

Web Application and User Interface

A lightweight web interface enables structured data submission and clear visualization of recommendation outputs, including calibrated probability scores and confidence metrics. The system incorporates secure authentication mechanisms and structured storage of patient-level records for future reference and auditability. This design prioritizes interpretability, traceability, and seamless integration within clinical workflow environments.

Conversational Assistant

An integrated large language model (LLM) component provides supplementary explanatory context for model predictions. While not directly influencing classification outputs, the LLM enhances usability by translating structured outputs into clinician-readable reasoning summaries.

Database Layer

A persistent storage layer supports logging of inputs, predictions, and confidence distributions, enabling reproducibility, auditability, and potential future calibration analysis.

Collectively, these components support practical deployment while preserving separation between the research-centered predictive model and auxiliary system features.

V. CONCLUSION

This work introduced *TriAID*, a structured decision-support framework for automated chest radiograph follow-up recommendation. By combining multi-label probabilistic modeling

with contextual large language model augmentation, the system bridges radiographic interpretation and actionable clinical guidance.

TriAID is designed to support physician decision making and enhance patient outcomes by providing structured, interpretable follow up recommendations.

Its modular architecture enables integration with scheduling systems and multidisciplinary communication workflows, supporting efficient coordination across care teams. Future development will focus on broader validation, improved calibration, and responsible deployment.

Furthermore, additional studies using de-identified images and reports from medical institutions would be of benefit to further assessing the model, along with responsible deployment and testing of the solution in a clinical setting.

Overall, TriAID represents a scalable and clinically aligned intelligent support tool, advancing artificial intelligence toward structured, interpretable, and workflow-compatible healthcare decision support.

VI. Acknowledgements

We would like to thank MedRAX open sourcing such a capable model for chest x-ray image classification, among other implementations the MedRAX team accomplished. We would also like to thank QMind for the opportunity to work on this project, and building the necessary frameworks to organize and continually support their numerous projects and members. Finally, we would like to thank our undisclosed radiology contact for their continued feedback and help in ensuring our solution was feasible and useful.

References

[1] K. L. Humphrey, M. D. Gilman, B. P. Little, E. F. Halpern, G. F. Abbott, J.-A. O. Shepard, and C. C. Wu, "Radiographic follow-up of suspected pneumonia," *Journal of Thoracic Imaging*, vol. 28, pp. 240–243, 12 2012. [Online]. Available: https://journals.lww.com/thoracicimaging/fulltext/2013/07000/radiographic_follow_up_of_suspected_pneumonia_.5.aspx

[2] B. P. Little, M. D. Gilman, K. L. Humphrey, T. K. Alkasab, F. K. Gibbons, J.-A. O. Shepard, and C. C. Wu, "Outcome of recommendations for radiographic follow-up of pneumonia on outpatient chest radiography," *American Journal of Roentgenology*, vol. 202, pp. 54–59, 01 2014.

[3] H. B. Harvey, M. D. Gilman, C. C. Wu, M. S. Cushing, E. F. Halpern, J. Zhao, P. V. Pandharipande, J.-A. O. Shepard, and T. K. Alkasab, "Diagnostic yield of recommendations for chest ct examination prompted by outpatient chest radiographic findings," *Radiology*, vol. 275, pp. 262–271, 12 2014.

[4] P. Xiao, X. Yu, S. M. Ha, A. Bani, A. Mintz, J. Wang, M. Elbanan, M. Mokkarala, G. Mattay, A. Nazeri, T. Kannampallil, A. M. Lai, V. R. Narra, D. S. Marcus, A. J. Bierhals, and A. Sotiras, "Large-scale evaluation of machine learning models in identifying follow-up recommendations in radiology reports," *Radiology*, vol. 317, 11 2025.

[5] A. Fallahpour, J. Ma, A. Munim, H. Lyu, and B. Wang, "Medrax: Medical reasoning agent for chest x-ray," arXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2502.02673v1

[6] S. R. Weingarten, B. Ermann, M. S. Riedinger, P. K. Shah, and A. G. Ellrodt, "Selecting the best triage rule for patients hospitalized with chest pain," *The American Journal of Medicine*, vol. 87, pp. 494–500, 12 2004. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0002934389806035

[7] bowang lab, "Github - bowang-lab/medrax: Medrax: Medical reasoning agent for chest x-ray - icml 2025," GitHub, 2025. [Online]. Available: https://github.com/bowang-lab/MedRAX

[8] X. Zhang, J. N. Acosta, J. Miller, O. Huang, and P. Rajpurkar, "Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports," arXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2505.00228

[9] A. Guo and T. Yang, "Research and improvement of feature words weight based on tfidf algorithm," IEEE Xplore, p. 415–419, 05 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7560393?casa_token=T8aedn2kqW4AAAAA:z5lSxG3OrrsJ6Uf_qx3Vduto7Rx8VmBUn3GzWZgoBicOjd6RQbZdKSETzZ2A0WUmgAqFU8qK

[10] I. O. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and..." ResearchGate, 02 2022. [Online]. Available: https://www.researchgate.net/publication/358284895_IDEAL_DATASET_SPLITTING_RATIOS_IN_MACHINE_LEARNING_ALGORITHMS_GENERAL_CONCERNS_FOR_DATA_SCIENTISTS_AND_DATA_ANALYSTS

[11] S.-L. , "sklearn.multiclass.onevsrestclassifier — scikit-learn 0.23.1 documentation," scikit-learn.org, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[12] Scikit-Learn, "sklearn.ensemble.randomforestclassifier — scikit-learn 0.20.3 documentation," Scikit-learn.org, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[13] H. Face, "sentence-transformers/all-minilm-l6-v2 · hugging face," huggingface.co, 08 2021. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[14] A. C. o. R. , "Acr.org home," Acr.org, 2018. [Online]. Available: https://www.acr.org/

[15] DeepSeek, "Deepseek," www.deepseek.com, 2025. [Online]. Available: https://www.deepseek.com/

[16] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize f1 score," *arXiv:1402.1892 [cs, stat]*, 05 2014. [Online]. Available: https://arxiv.org/abs/1402.1892