MediAnnote: A Framework for Collaborative Annotation and Retrieval of Chest X-Rays Using the Radiology Gamuts Ontology

Mona Alshlowi
Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
mdhaifallahalshlowi@stu.kau.edu.sa (corresponding author)

Samar Alkhuraij
Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
salkhuraiji@kau.edu.sa

Hajar Alharbi
Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
hmsalharbi@kau.edu.sa

Received: 8 November 2025 | Revised: 26 November 2025 | Accepted: 13 December 2025

 DOI: https://doi.org/10.48084/etasr.16118

ABSTRACT

Medical imaging is a cornerstone of healthcare, supporting disease diagnosis, treatment planning, and clinical research. The growing volume of medical imaging data has made image annotation an increasingly complex yet essential step in producing reliable, high-quality labeled datasets for diagnostic, educational, and research purposes. This work presents MediAnnote, an ontology-based framework for medical image annotation and retrieval that combines a deep learning component for pre-annotation with the Radiology Gamuts Ontology (RGO) within a collaborative environment. The framework enables image retrieval along with the corresponding findings and causes, enhancing both educational and clinical applications. MediAnnote outperformed existing annotation systems in a qualitative comparison incorporating all essential components. An experimental study involving three radiologists and the NIH Chest X-ray dataset showed that the model achieved a higher accuracy in disease prediction, with an F1-score of 0.54, an AUC of 0.75, a precision of 0.54, and a recall of 0.53, compared to individual radiologists. In addition, integrating a human-in-the-loop approach improved the precision of abnormality localization. The post-task survey showed high user satisfaction, with an overall mean score of 3.94 out of 5.

Keywords-ontology; medical image annotation; semantic image retrieval; chest x-ray; semantic web; human-in-the-loop; deep learning; radiology AI

I. INTRODUCTION

Chest diseases remain a global health issue. In resource-limited regions, conditions such as tuberculosis, pneumonia, interstitial lung disease, and lung cancer still strain healthcare systems [1, 2]. Chest X-ray is the most widely used and cost-effective tool among diagnostic modalities, valued for its accessibility and rapid image acquisition in both advanced hospitals and primary care centers [3]. However, the growing volume of imaging data and the increasing demand for early detection have significantly increased the workload of radiologists, who must interpret and annotate numerous chest images daily [4].

Annotation, to identify key structures and diseases in medical images, is essential in developing reliable AI models for radiology, but remains time-consuming and resource-intensive [5, 6]. Since it requires expert knowledge and close attention to subtle clinical details, annotation is often a significant bottleneck in medical AI development, as radiologists have limited time, but their input is vital for diagnostic accuracy [7, 8]. Annotation inconsistencies often stem from variations in radiologists' interpretations and terminology, resulting in ambiguous or conflicting labels that reduce data sharing and model generalization [9, 10].

Ontologies such as RadLex [11], SNOMED-CT [12], and Radiology Gamuts Ontology (RGO) [13] offer standardized vocabularies that describe the relationships between imaging findings and possible diagnoses. The RGO organizes radiological "range results" along with their related differential diagnoses, providing a comprehensive knowledge framework to support diagnostic reasoning [13]. Incorporating these ontologies into the image annotation process improves semantic precision, promotes interoperability, and increases the reusability of annotated data in both clinical and research environments.

Recent advances in deep learning have transformed medical imaging by allowing rapid and accurate detection, classification, and segmentation of chest diseases [14, 15]. Such models perform well in biomedical image classification [16], but their reliance on unstructured annotations limits semantic understanding and the reuse of datasets. A systematic mapping study on deep learning-based ontology learning showed that most research remains focused on textual data, with little progress towards integrating ontologies into image-based or multimodal frameworks [17].

Relying on automated systems does not guarantee diagnostic reliability because AI models are prone to errors, data bias, and limited transparency [18-20]. As a result, there has been growing interest in hybrid approaches that integrate human expertise through Human-In-The-Loop (HITL) approaches, involving radiologists who review, adjust, and verify AI-generated outputs [21, 22]. This collaborative approach improves annotation quality, supports model refinement through active learning, and maintains the essential role of clinical expertise in diagnostic decision-making.

At the tool level, a variety of open-source and commercial web-based and standalone image annotation tools have been developed to aid data labeling across domains such as pathology, radiology, and computer vision. This study limited its scope to open-source web-based platforms. The majority of available tools range widely in both functionality and interface design. Some are elementary polygonal markup tools that provide basic functionality, while others offer a more complex, AI-assisted, and collaborative environment.

Despite advances in AI-powered annotation and ontology-based structuring, a critical research gap remains, as current systems rarely integrate all three of AI pre-annotation, ontology-driven semantic labeling, and human validation into a single framework. Most existing tools focus only on automation or ontology management. Therefore, there is an urgent need for an ontology-based annotation platform, integrated with AI, to automatically generate initial labels, import standardized terms from ontologies such as RGO, and allow radiologists to validate and improve results.

LabelMe [23] and ImageTagger [24] were among the first systems to establish web-based image annotation standards. Although these systems provided straightforward manual functions for image annotation and the creation of computer vision training datasets, they neither met the standards specific to medical imaging nor supported the inclusion of the semantic structures needed for clinical datasets.

More recent web-based DICOM viewer systems, such as DWV [25] and OHIF [26], offered new, simplified ways to view and annotate medical images in web browsers. Both systems increased accessibility and interoperability but lacked label management, ontology integration, and multi-annotator collaboration.

Recent years have seen a flurry of advanced open source frameworks for medical image annotation, including VinDrLab [27], MONAI Label [28], and EXACT [29]. VinDrLab provides web-based multi-annotator annotation workflows for radiology datasets that support DICOM images. MONAI Label includes deep learning models for real-time AI pre-annotation and interactive correction of annotations. EXACT provides algorithm-aided annotation, version control, and multi-user collaboration for extensive pathology and radiology datasets. At the same time, several platforms, such as Cytomine and Quick Annotator [30, 31], have focused on digital pathology and have included support for large image format (WSI) and team collaboration. However, they lack ontology-driven labeling and structured semantic validation. There are also general-purpose platforms, e.g., CVAT [32] and Label Studio [33], that integrate AI with different image types, but these require more customization for medical use and do not include human involvement and ontology integration.

Finally, newer experimental systems, including LOST [34], ImgLab [35], and MedTAG [36], emphasize active learning, lightweight web deployment, and text-image joint annotation. Although these systems improve annotation speed, none incorporate ontology-based consistency mechanisms or semantic search capabilities.

Existing platforms like MONAI Label, EXACT, and VinDr Lab support AI-assisted annotation and collaborative workflows, but do not offer a single integrated solution that combines ontology-based semantic labeling, AI Assistant, HITL correction, and semantic retrieval, despite advances in medical image labeling capabilities. There is clearly a need for a unified solution to address these gaps and improve the overall efficiency and usability of the annotation process. The proposed MediAnnote integrates these elements into a single, coherent framework, combining AI pre-annotation with structured label validation using the RGO, supporting hierarchical reasoning, and enabling semantic search across annotated images. To the best of our knowledge, no existing open-source system provides this full end-to-end capability. This integration allows MediAnnote to deliver a more consistent, interpretable, and clinically aligned annotation experience for radiologists.

To address these challenges, this paper introduces an ontology-based framework, called MediAnnote, using the DACNET model [37] to generate preliminary labels that are then validated and refined by radiologists, including semantic retrieval to allow a structured search. The tool is designed for applications in diagnosis, education, and research. To ensure that the framework integrates all main components, interviews were conducted with three experienced radiologists, each with 7 years of field experience, to identify the key elements of a practical framework. Finally, quantitative and qualitative evaluations were performed to assess the proposed tool.

The main contributions of this study are:

- The proposal and implementation of an image annotation and retrieval framework that integrates ontology, AI assistance, HITL mechanisms, and semantic retrieval.
- Quantitative and qualitative evaluation of the performance and usability of the MediAnnote framework.
- Integration of RGO into the framework to ensure semantic consistency.
- Mapping of the 14 NIH Chest X-ray classes to their corresponding RGO terms.

II. METHODS

A. Identifying the Key Framework Components

Interviews were conducted with three radiologists, with seven years of experience each, to identify the key components of the MediAnnote annotation framework.

1) Data Collection

The primary objective of the interviews was to determine the essential components and requirements for developing a robust medical image annotation framework that supports and assists healthcare professionals, particularly radiologists and other physicians who interpret medical imaging. The interview used a combination of open and multiple-choice questions designed to address challenges in medical image annotation and assess the value of controlled vocabularies, automated AI assistance, and collaborative features.

2) Participant Selection

Radiologists were selected based on their experience in diagnosing lung diseases using X-rays and their familiarity with existing annotation tools. All three participants had at least seven years of professional practice with chest radiograph interpretation.

3) Interview Questions

The following are examples of the interview questions:

- What tools or software do you currently use for annotating and labeling different imaging modalities?
- What workflow or methods do you follow during the annotation process?
- Which features of existing annotation systems do you find most valuable or practical?
- What limitations or drawbacks have you experienced when using current annotation platforms?
- What challenges do you encounter when performing medical image annotation with the available tools and systems?

4) Data Analysis

All interviews were conducted online and lasted approximately 30 minutes each. A thematic analysis approach was applied to interpret the data and identify recurring needs, challenges, and preferences across participants.

a) Tool Requirements

From the interviews, several key features emerged as essential to the tool's design:

- Search and Retrieval: Efficient mechanisms to search and retrieve images.
- Collaboration: Features that enable teamwork.
- Ontology/Controlled Vocabulary: Use of standardized medical terminologies.
- AI Assistance: Support from AI models for preliminary annotations and intelligent suggestions.

b) Preferred System Specifications

The system should meet the following specifications:

- Remote Accessibility: Usable from any location.
- Open-Source and Free: Accessible without cost to ensure widespread adoption.
- User-Friendly Interface: Designed for ease of use, even for non-technical users.
- The need for ontology-based annotation to reduce errors and ensure consistency.
- A preference for AI-assisted tools to speed up the annotation process.
- The importance of collaboration features for multi-user environments.
- The value of educational integration for training and mentoring medical students.

B. Evaluation

The MediAnnote framework was evaluated through three assessments: a comparative analysis, a quantitative evaluation of image annotation performance, and a post-task survey.

III. PROPOSED FRAMEWORK

The proposed framework, illustrated in Figure 1, comprises two core modules—the annotation module and the retrieval module. The annotation module includes AI pre-annotation, ontology integration, HITL, and collaboration features. The retrieval module leverages ontology integration for semantic search. The system was designed based on the requirements and challenges identified during the interviews. These components are recognized as essential for implementing a practical collaborative semantic medical image annotation and retrieval tool for medical practitioners and students.

A. Framework Structure

- AI Pre-annotation: The AI-assisted annotation in MediAnnote is powered by DACNet, a pre-trained deep learning model for chest X-ray analysis. DACNet was selected due to its highest reported average AUC in prior studies [37]. The model was used off-the-shelf, and the original architecture, training process, hyperparameters, dataset splits, and pre-processing steps are detailed in the original publication. In the proposed framework, DACNet

generates preliminary annotations with editable bounding boxes, and Grad-CAM is applied for localization, producing heatmaps that assist radiologists in an HITL workflow.

- Ontology Integration: All findings' labels are imported from the RGO. The mapping between the NIH ChestX-ray14 dataset labels [38] and their corresponding RGO terms ensures consistent terminology across AI-generated and expert-validated annotations. The term "Pulmonary Nodule" was added based on expert consensus and referenced from Radlex [39] and the Human Phenotype Ontology [40]. In the RGO, "pulmonary consolidation" appears as a synonym for "alveolar lung disease" (ID: 14175), as shown in Table I.
- HITL and Collaboration: Radiologists validate AI suggestions through a two-level process comprising annotator refinement followed by reviewer approval. Collaborative features allow multiple experts to work on the same dataset, enabling discussion, consensus-building, and high-quality annotations.
- Ontology-based Retrieval: Indexed RGO concepts allow semantic search to retrieve approved annotated images based on their findings, with their causes, and vice versa. This enables structured retrieval beyond simple keyword searches, supporting educational and research use cases.
- Storage: Handles medical image data, including storage and retrieval of both unlabeled and annotated images, as well as associated ontology mappings, user profiles, and project metadata.

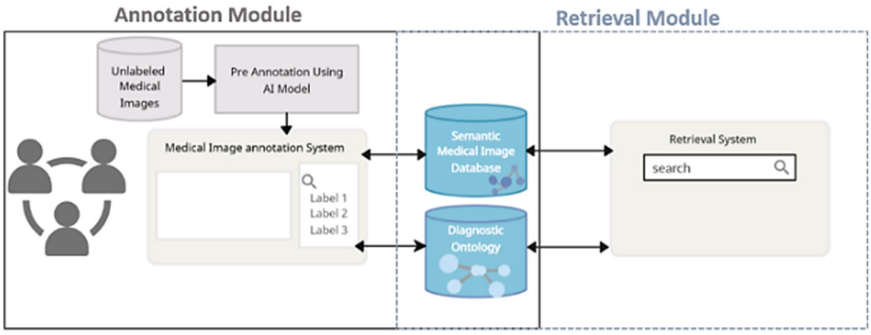

Fig. 1. Framework of collaborative medical image annotation and retrieval.

TABLE I. MAPPING NIH CHESTX-RAY14 DATASET LABELS TO RGO TERMS

NIH_Label	Preferred_RGO_Term	Confirmed_RGO_ID
Atelectasis	Atelectasis	rgo:15792
Cardiomegaly	Cardiomegaly	rgo:22537
Effusion	Pleural effusion	rgo:03688
Infiltration	Pulmonary opacity	rgo:32683
Mass	Pulmonary mass	rgo:03705
Nodule	Pulmonary nodule	Added
Pneumonia	Pneumonia	rgo:03672
Pneumothorax	Pneumothorax	rgo:23339
Consolidation	Pulmonary consolidation	Alt: rgo:14175
Edema	Pulmonary edema	rgo:14286
Emphysema	Emphysema	rgo:03584
Fibrosis	Pulmonary Fibrosis	rgo:21475
Pleural Thickening	Pleural thickening	rgo:15999
Hernia	Hernia	rgo:24748

B. Technical Implementation

The tool was developed using the following technologies:

- Front end: A web-based interface built with React.js for ease of use and accessibility.
- Back end: A Python-based server leveraging Django.
- AI model: DACNET, a pre-trained deep learning model for annotation suggestions; the Grad-CAM method is used for localization, as illustrated in Figure 2.

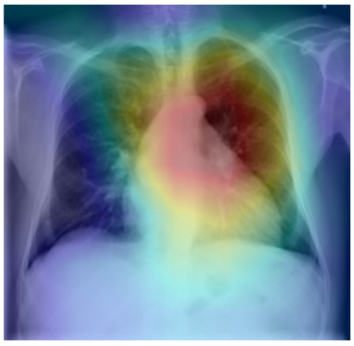

Fig. 2. Grad-CAM method for localization.

- Database: SQLite for storing projects, users, images, annotations, and ontology structures.
- Ontology: The RGO is integrated through the OWL API and Elasticsearch (Docker) for efficient indexing and semantic retrieval. NIH synonyms API is leveraged for synonym expansion in both annotation and retrieval.

IV. SYSTEM DESCRIPTION

The system supports three distinct roles: Admin, Annotator, and Reviewer.

A. Project Administration

Administrators are given the following functions, as shown in Figure 3.

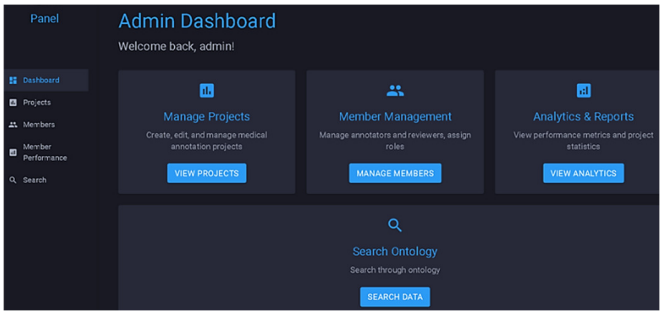

Fig. 3. Admin dashboard.

1) *Project Management*

Admins can create/delete new projects, upload image datasets, and import labels, as shown in Figure 4.

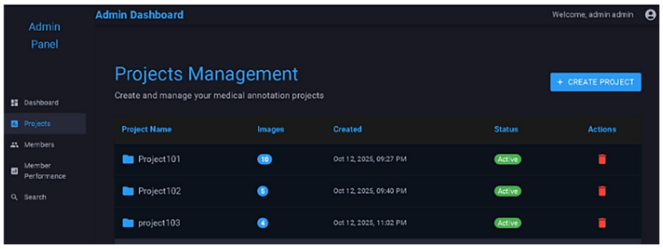

Fig. 4. Project Management.

2) *User Management*

Admins can add/delete team members to/from a project and assign roles (annotator or reviewer) as shown in Figure 5.

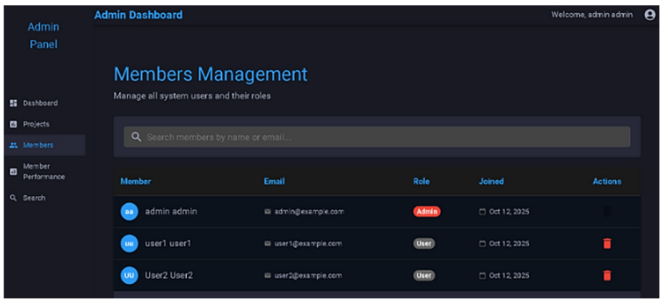

Fig. 5. Members' management.

3) *Statistics and Reporting*

The system tracks annotator performance as shown in Figure 6, displaying metrics such as:

- Number of labeled images.
- Time spent per image.
- Review acceptance rate, which is the number of images returned by reviewers for modification divided by the number of labels.

These statistics can be exported as an Excel file for further analysis.

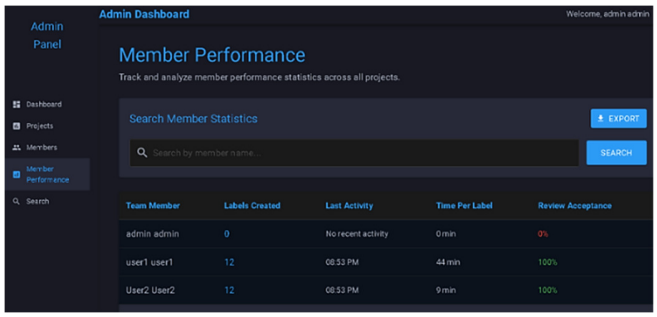

Fig. 6. Members' performance.

B. *Annotation Workflow*

1) *Project Setup*

After an Admin sets up the project, uploads datasets, and assigns members, the annotation process begins as shown in Figure 7.

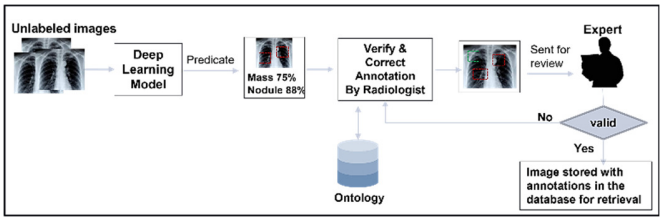

Fig. 7. Annotation process.

2) *Annotator Role*

The image is pre-annotated by the DACNET model. Annotators can modify or remove the AI-predictions and their bounding boxes. They can also add labels using bounding box tools, choosing labels from a predefined list imported from the ontology, as shown in Figure 8.

- Annotators can choose to display or hide the AI-generated predictions description, as shown in Figure 9.
- Once labeling is complete, the image is submitted for review. The image statuses are:
  - Unprocessed
  - For Review
  - Needs Modification
  - Approved

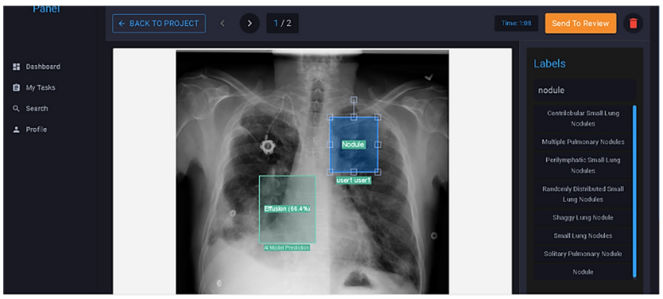

Fig. 8. Annotation interface.

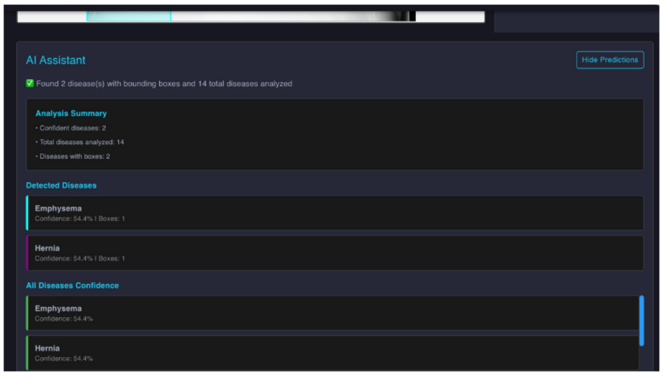

Fig. 9. AI predictions' description.

3) Reviewer Role

- Reviewers evaluate submitted annotations, either approving the image, which stores it in the database for retrieval, or marking it as "Needs Modification," which returns it to the annotator for corrections, as shown in Figure 10.
- Following the modification, the annotator resubmits the image for final approval.

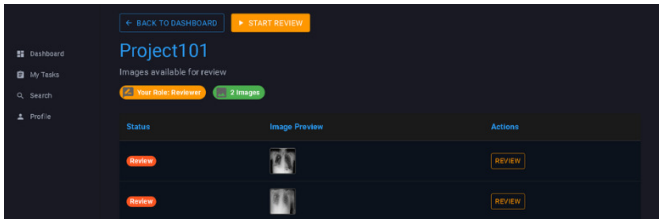

Fig. 10. Review interface.

C. Retrieval Workflow

All roles (Admin, Annotator, and Reviewer) have access to the search function, allowing images to be retrieved from the database and filtered.

1) Identification of an Imaging Finding to Search for

The process begins with the identification of a single key imaging finding in a patient's medical image. A radiologist typically makes this observation during the annotation.

2) Querying RGO for Causal Relationships Once an Imaging Finding is Identified

The medical image retrieval system queries the RGO to explore causal relationships for the selected finding.

3) Retrieval of Stored Images with Potential Causes and Differential Diagnosis

Images are retrieved from the database, along with a comprehensive list of potential causes from RGO. For example, suppose that a radiologist reviews a chest X-ray and observes pleural thickening, which is the key imaging finding identified. Then, the system queries RGO for conditions that may cause or may be caused by pleural thickening and presents the radiologist with the images and a list of causes and their converse, as shown in Figure 11.

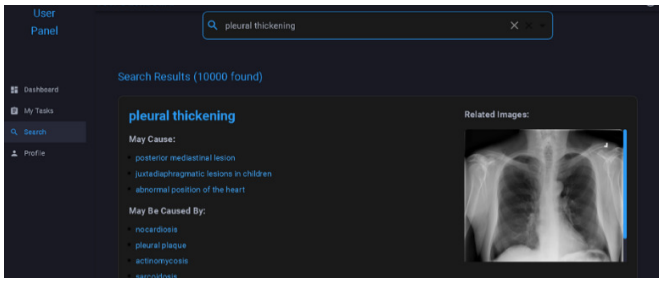

Fig. 11. Retrieval result using Gamut ontology.

V. EXPERIMENTAL SETUP

A. Dataset Description

The publicly available NIH ChestX-ray14 dataset was employed, which contains 112,120 frontal chest radiographs from 30,805 unique patients, annotated with 14 common thoracic pathologies, including atelectasis, effusion, infiltration, pneumonia, and cardiomegaly [38], as shown in Figure 12. This publicly available and fully de-identified dataset can be accessed through the official NIH Box repository [41] or through the Kaggle-hosted mirror [42]. All radiologist annotations were performed on existing images, and no new patient data were collected; therefore, Institutional Review Board (IRB) approval was not required.

This dataset provides image-level disease labels but lacks bounding-box ground truth localization for some classes. The lack of bounding-box ground truth localization restricts the use of fully automated evaluation for the localization of findings, necessitating human involvement in refining annotations. For this study, a subset of 51 images with 60 labels, including some with multiple labels, was randomly selected for AI-assisted annotation and manually validated by three radiologists with 7, 7, and 5 years of experience, respectively.

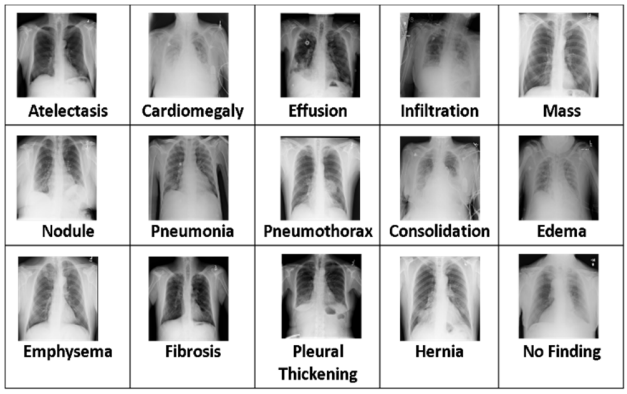

Fig. 12. 14 classes of the NIH chest X-Ray dataset.

B. AI Model and Ontology Integration

The annotation workflow incorporated a DACNET convolutional neural network trained for multilabel disease classification [37]. Predicted labels were automatically mapped to corresponding RGO terms to ensure semantic consistency across annotators and institutions. Moreover, all general labels from the AI model that are not included in the RGO are added

to it. This ontology layer provided structured terminology and hierarchical relationships between radiological findings, enabling precise label standardization and retrieval, as shown in Figure 13.

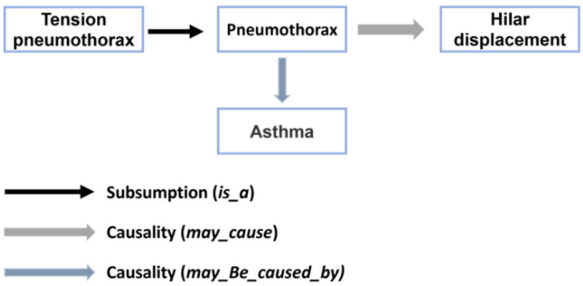

Fig. 13. Gamut Ontology Relation.

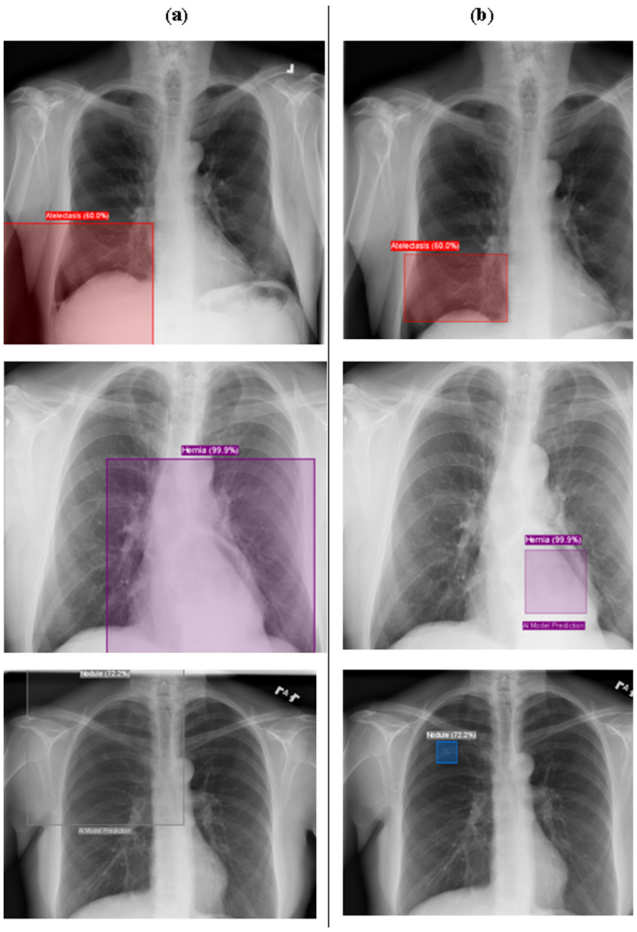

Fig. 14. Sample of test set: (a) bounding box by AI; (b) Adjusted bounding box by radiologists.

C. Experimental Procedure

Three radiologists participated in the user evaluation. One radiologist was previously involved in the interview phase, whereas the remaining two were independent participants, each completing both annotation and post-survey tasks. For the annotation task, each participant corrected AI-generated labels, added ontology terms when necessary, adjusted the bounding box location, and submitted validated annotations. Figure 14 illustrates how human intervention improves lesion localization accuracy by comparing AI-predicted bounding boxes with radiologist-corrected annotations.

D. Evaluation Metrics

To evaluate diagnostic performance, AUC, Precision, Recall, and F1-score were calculated for both AI and human annotators. The consistency among radiologists was assessed using Cohen's κ for pairwise agreement. Finally, user satisfaction and perceived usability were evaluated through a post-task Likert-scale online survey implemented in Google Forms, comprising 20 items grouped into five dimensions: usability, annotation efficiency, retrieval effectiveness, collaboration & workflow, and ontology utility & AI suggestion.

VI. RESULTS AND DISCUSSION

A. Comparative Evaluation of Existing Platforms

A comparative review of 12 existing web-based, open-source medical image annotation platforms was conducted. Table II summarizes the evaluation of their main components.

TABLE II. COMPARATIVE ANALYSIS OF WEB-BASED OPEN-SOURCE MEDICAL IMAGE ANNOTATION PLATFORMS

Platform	Domain	AI-Assisted Annotation	Ontology Integration	HITL Workflow	Retrieval /Search feature
LabelMe	General	No	No	No	No
DWV	Radiology	No	No	No	No
OHIF Viewer	Radiology	Plugin AI	No	Manual review	Metadata only
Slim Viewer	Pathology/ Radiology	Limited	No	No	No
VinDr Lab	Radiology	Basic	No	Reviewer role	No
CVAT	General	Yes	No	Review UI	No
Cytomine	Pathology	Limited	No	Collaborative	Metadata search
Quick Annotator	Pathology	DL iterative	No	Correction loop	No
MONAI Label	Medical (3D/CT/ MRI)	Active learning	No	Yes	No
Label Studio (CE)	Multi-domain	Plugin-based	No	Review UI	No
StudierFenster	Radiology (CT/MR)	Limited	No	Validation only	No
EXACT	Radiology/ Pathology	Algorithm-aided	No	Versioned review	Basic image index
MedTAG	Biomedical text/image	Suggestion API	Light ontology	Collaborative	Text search
PIMIP	Pathology	ML plugin	No	Multi-annotator	No
ImgLab	General	Limited	No	No	No
ImageTagger	General	No	No	Limited	No
LOST	General	Semi-automated	No	Partial	No
MediAnnot	Radiology	AI-assisted	Yes	Yes	Yes, semantic search

Most existing tools focus on manual labeling or AI-assisted segmentation, with limited integration of semantic ontologies to ensure consistency of annotation. Furthermore, while platforms such as MONAI Label support AI-driven annotations, they lack the ontology-based integration and semantic retrieval features that are essential for clinical interoperability and dataset reuse. The proposed framework, MediAnnote, introduces a unified tool that integrates an AI pre-annotation DACNET model, ontology guided via the GRO, a HITL validation process, and retrieval. Thus, the comparison highlights the system's novelty in combining AI, ontology, and human expertise within a single web-based platform.

B. Quantitative Evaluation of Image Annotation Performance

Tables III and IV show that the DacNet model achieved the highest overall performance (AUC=0.75, F1=0.54), outperforming Radiologist 1 (AUC=0.63), Radiologist 2 (AUC=0.70), and Radiologist 3 (AUC=0.65). Radiologist 2 showed the best human performance, particularly in identifying cardiomegaly (0.99) and images that did not warrant a finding (0.82). The AI model excelled in cardiomegaly, edema, and atelectasis, but both AI and radiologists achieved lower accuracy on pneumonia and consolidation. These results show that AI provides strong pre-annotations, while expert review ensures diagnostic precision and clinical consistency.

TABLE III. PER-DISEASE AUC COMPARISON BETWEEN THE DACNET MODEL AND THREE RADIOLOGISTS ON THE NIH CHESTX-RAY14 DATASET

Disease	AI_AUC	R1_AUC	R2_AUC	R3_AUC
Cardiomegaly	0.9787	0.9574	0.9894	0.9149
Edema	0.9468	0.7181	0.6144	0.8723
Hernia	0.8644	0.6250	0.7394	0.6144
Mass	0.8644	0.7500	0.8644	0.8644
Atelectasis	0.8537	0.8431	0.8644	0.7234
Fibrosis	0.8431	0.5000	0.4894	0.5000
Infiltration	0.8431	0.4894	0.7500	0.5000
Nodule	0.7500	0.6250	0.6144	0.6144
Pneumothorax	0.7394	0.6144	0.7181	0.6144
Effusion	0.7181	0.7074	0.7394	0.7793
Emphysema	0.7074	0.7394	0.7394	0.6037
Pleural Thickening	0.6144	0.5000	0.6250	0.5000
Consolidation	0.5000	0.4574	0.4787	0.6037
No Findings	0.5000	0.4761	0.8191	0.5000
Pneumonia	0.4894	0.5000	0.5000	0.5000
Average	0.7475	0.6335	0.7030	0.6470

TABLE IV. OVERALL PERFORMANCE OF THE DACNET MODEL AND THREE RADIOLOGISTS ACROSS ALL DISEASES

Source	Precision	Recall	F1-score	AUC
DACNET Model	0.54	0.53	0.54	0.75
Radiologist 1	0.35	0.32	0.33	0.63
Radiologist 2	0.47	0.45	0.46	0.70
Radiologist 3	0.27	0.38	0.32	0.65

C. Usability and Post-Task Survey Results

The results of the post-task survey in Table IV summarize user ratings across five evaluation dimensions, measured using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) The survey included 20 items in total: Usability and Learnability (4 items), Annotation Effectiveness (3 items), Retrieval Effectiveness (3 items), Collaboration and Workflow (5 items), and AI Suggestion and Ontology Utility (4 items). Participants reported positive experiences with the framework, reflected by an overall mean score of 3.94±0.40 and 78% agreement for ratings of 4 or higher.

Within individual dimensions, users strongly approved the annotation workflow, as demonstrated by high ratings for Annotation Effectiveness (4.56±0.51) and Usability and Learnability (4.17±0.29). Collaboration and Retrieval also received favorable responses, with mean values of 3.87±0.42 and 3.78±0.51 indicating general satisfaction with teamwork features, version control, sharing, and search capabilities. The Semantic Consistency dimension (3.33±0.29) showed less agreement, suggesting that improvements in ontology-based suggestions and semantic retrieval could further enhance user experience.

TABLE V. DESCRIPTIVE ANALYSIS OF USER EVALUATIONS

Dimension	Mean ± SD	Agreement (≥4)	Interpretation
Usability and Learnability	4.17 ± 0.29	100%	Excellent usability and simple to learn
Annotation Effectiveness	4.56 ± 0.51	100%	Highly effective annotation tools and labeling workflow
Retrieval Effectiveness	3.78 ± 0.51	67%	Moderate satisfaction with retrieval precision
Collaboration & Workflow	3.87 ± 0.42	80%	Good collaboration, version control, and sharing support
AI Suggestion and Ontology Utility	3.33 ± 0.29	42%	Needs improvement in AI suggestions
Overall Mean	3.94 ± 0.40	78%	High overall acceptance and usability

D. Impact on Diagnostic Workflows

The ontology-driven labeling and semantic retrieval capabilities of MediAnnote provide structured and standardized annotations that reduce inconsistencies and ambiguity in medical image labeling. By mapping predicted findings such as Atelectasis, Cardiomegaly, Effusion, Infiltration, Pneumonia, Nodule, Mass, Pneumothorax, Consolidation, and Edema to RGO, the system allows radiologists to access related images and associated diagnoses efficiently. Each disease in the ontology includes two types of relationships, may_cause and may_be_caused_by (Figure 11), which facilitates diagnostic reasoning by allowing users to iteratively explore causal and associated radiology. The NIH synonyms API is leveraged to expand label coverage and improve semantic search, enabling users to retrieve relevant images and findings using both direct disease terms and their synonyms.

Overall, the survey findings confirm that the proposed framework is usable, practical, and well accepted by participants, while also identifying potential refinements in semantic reasoning.

E. Retrieval Performance and Limitations

Semantic retrieval in MediAnnote relies on RGO to return images annotated with the queried term or its synonyms. For each clinical query (e.g., "Atelectasis," "Cardiomegaly"), the

system retrieves all images linked to the corresponding ontology concept, ensuring that every returned result is semantically valid. In this evaluation, each queried finding was associated with four images in the dataset, and MediAnnote accurately retrieved all four in every case, resulting in Precision@K of 1.0 and Recall@K of 1.0. Because the retrieval mechanism is ontology-driven and rule-based, the system does not perform similarity ranking; instead, it guarantees complete and accurate coverage of images mapped to the queried concept. While these results demonstrate the correctness and reliability of the retrieval component within a controlled dataset, future work will include scaling the evaluation to larger collections, introducing similarity-based ranking, and optimizing retrieval performance under real clinical workloads.

VII. DISCUSSION

MediAnnote was compared to existing annotation frameworks, as shown in Table II. The proposed system outperformed all other tools in areas such as integration of AI with ontology, two-level human validation, collaboration support, and retrieval functionality. This combination provides a standardized environment for collaborative medical image annotation.

The experimental findings corroborate and extend several trends identified in prior research on ontology-driven annotation, HITL systems, and AI-assisted image analysis. Quantitatively, the proposed framework achieved an average AUC of 0.75, consistent with previously reported performance ranges for deep learning models applied to thoracic disease detection. However, it is essential to note that these results are based on a small subset of 51 images with 60 labels. While the subset was sufficient for assessing usability, HITL interaction, and ontology integration, it limits the statistical strength of the reported AUC and F1-score values. Therefore, these findings should be interpreted as outcomes of a feasibility or pilot study rather than definitive performance benchmarks. Unlike conventional AI-based approaches that rely on unstructured labeling, the integration with RGO introduced a layer of semantic validation, addressing the inconsistencies observed in earlier studies.

This study builds on [37], which reproduced and improved the CheXNet model for chest X-ray classification, achieving strong accuracy (AUC≈0.85) on the NIH ChestX-ray14 dataset. The proposed framework used the same model for AI-based pre-annotation, but it was extended with an ontology-guided, expert-validated process. The model's predictions were mapped to standardized terms from the RGO and reviewed by radiologists to ensure consistency and clinical relevance. Unlike the fully automated approach in [37], the proposed system integrates human validation and ontology reasoning to produce more accurate, consistent, and interpretable image annotations.

The AI-assisted model achieved strong classification performance (AUC=0.75) on the dataset, outperforming the average radiologist's accuracy, and emphasized its role as a pre-annotation aid. The annotation results revealed several important insights. Radiologists unanimously identified findings that contradicted the NIH ChestX-ray14 ground-truth labels in 3 out of 51 images (5.88%). This discrepancy demonstrates that the weakly supervised labels in the original dataset do not always align with expert diagnostic consensus. The result highlights the value of the HITL mechanism in detecting mislabeled cases and improving the reliability of annotated datasets.

Since the NIH ChestX-ray14 dataset does not include bounding-box ground truth for the selected images, standard quantitative localization metrics such as IoU or mAP could not be computed. Therefore, a qualitative comparison was made between the three radiologists and Grad-CAM heatmaps through a structured visual inspection. Radiologists' localization of the findings was more specific and anatomically precise than the bounding box generated by the AI model. Across the evaluated images, all three radiologists independently noted that AI-generated localizations tended to be diffuse and span broader regions. In contrast, their own annotations were more focused on anatomically relevant structures. Although AI correctly identified pathological areas in most cases, its predicted heatmaps tended to be broader. This assessment is purely based on qualitative rather than quantitative data. The three radiologists, working independently, agree that the AI heatmaps require refinement and that human validation is critical for producing clinically meaningful localization.

User feedback from the post-task survey highlights that the MediAnnote framework successfully balances functionality, usability, and semantic integration. The participants described the system as intuitive and easy to learn, reflected in strong scores for Usability and Learnability (4.17±0.29). The Annotation Effectiveness rating (4.56±0.51) further demonstrates that users found the AI-assisted ontology-guided labeling workflow both efficient and reliable for diagnostic annotation. Moderate scores for Collaboration (3.87±0.42) and Retrieval (3.78±0.51) indicate that users valued the teamwork and image search features but identified opportunities to enhance coordination and result filtering. The comparatively lower score in AI Suggestion and Ontology Utility (3.33±0.29) suggests that although the users recognized the value of automated label prediction and ontology alignment, they desired greater accuracy and more intuitive semantic search capabilities.

Although radiologists expressed overall support for the integration of AI assistance, they noted several instances in which AI-generated labels did not align with their diagnostic interpretations. For example, the model prediction "Infiltration" was mapped to the ontology term "Pulmonary opacity," which at times resulted in ambiguity or misinterpretation. This reflects broader challenges associated with semantic alignment between model outputs and the RGO. Such inconsistencies may undermine user trust and underscore the need for clearer, clinically intuitive terminology mappings and refinements to both ontology-driven retrieval and AI suggestion mechanisms.

Overall, these findings demonstrate that the proposed framework achieves high annotation accuracy and usability and bridges human diagnostic precision with AI scalability. However, it should be noted that the survey involved only three

radiologists, which is a small sample size. Feedback provided valuable preliminary insights into system usability and workflow integration. Future work will include a larger, more diverse group of participants from multiple institutions and varying experience levels to enable a statistically robust, generalizable evaluation of the framework's usability and effectiveness.

Despite MediAnnote's promising capabilities, the experimental results should be interpreted with caution. The DACNet model used for AI pre-annotation was selected based on prior literature reporting one of the highest AUC values on the ChestX-ray14 dataset; however, in this feasibility study, it yielded an F1-score of 0.54. Several factors contributed to this lower performance. First, the radiologists noted that many NIH images were suboptimal, with low resolution and artifacts that reduced diagnostic confidence and annotation consistency. Second, certain conditions, particularly cardiothoracic findings such as Cardiomegaly, typically require multi-view imaging for accurate assessment. Finally, the small experimental dataset (51 images annotated by three radiologists) constrains statistical reliability.

VIII. CONCLUSION AND FUTURE WORK

Accurate and efficient medical image annotation is essential for diagnosis, education, and research. However, existing tools often lack key features, including ontology-based consistency, AI automation, searchability, collaboration capabilities, and educational applications. The proposed ontology-based framework integrates key features encompassing all of these concepts into a single tool. This study experimented on 51 images with 60 labels using the NIH Chest X-ray database, with assistance from a deep learning model and three radiologists. The findings show that having an HITL for image annotation greatly enhances annotation accuracy. To this end, adding RGO to the search process enables semantic search and, therefore, facilitates the use of the system to retrieve images with their relevant causes, thus making it particularly useful for educational and collaborative settings.

However, several significant limitations must be acknowledged. First, radiologists noted that the image quality in the NIH dataset was sometimes suboptimal, which affected diagnostic confidence and annotation consistency. Second, specific pathologies, such as Cardiomegaly, typically require multi-view imaging (e.g., PA and lateral views) for reliable assessment. This constraint inherently limits the human ability to annotate a specific disease. Third, using only 51 images and three radiologists reduces the statistical strength of the results and the generalizability of the findings.

The current implementation is limited to chest X-rays and English-only ontology mappings within a specific clinical domain. In future work, the proposed framework can be extended and validated across other clinical domains, datasets, AI models, and imaging modalities. Improving the AI model's localization capabilities will further enhance annotation precision. The system can also be expanded to incorporate additional and multilingual ontologies (e.g., English-Arabic), enabling richer inference and reasoning. Furthermore, future development will focus on large-scale retrieval, ranking strategies, and latency optimization. Finally, expanding the validation dataset, incorporating multi-view imaging, and involving a larger pool of radiologists will help support a more comprehensive assessment of the framework's robustness and generalizability.

REFERENCES

[1] T. Hailemariam *et al.*, "Chest X-ray predicts cases of pulmonary tuberculosis among women of reproductive age with acute respiratory symptoms: A multi-center cross-sectional study," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, vol. 32, Aug. 2023, Art. no. 100383, https://doi.org/10.1016/j.jctube.2023.100383.

[2] A. Q. Khan, S. Nayyar, S. Khalid, I. Taqi, and M. A. Khan, "Prevalence of Interstitial Lung Disease in Patients with Chronic Cough Taking Chest X-rays and CT Chest as Diagnostic Tools," *Journal of Islamabad Medical & Dental College*, vol. 13, no. 1, pp. 110–115, Apr. 2024, https://doi.org/10.35787/jimdc.v13i1.1073.

[3] A. Soni and A. Rai, "A systematic survey on deep learning techniques for chest disease detection using chest radiographs," *Journal of Current Science and Technology*, vol. 13, no. 2, pp. 267–295, July 2023, https://doi.org/10.59796/jcst.V13N2.2023.1744.

[4] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, July 2021, Art. no. 102062, https://doi.org/10.1016/j.media.2021.102062.

[5] A. Ryabtsev, R. Lederman, J. Sosna, and L. Joskowicz, "Streamlining the annotation process by radiologists of volumetric medical images with few-shot learning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 20, no. 9, pp. 1863–1873, June 2025, https://doi.org/10.1007/s11548-025-03457-3.

[6] M. Napravnik, F. Hržić, S. Tschauner, and I. Štajduhar, "Building RadiologyNET: an unsupervised approach to annotating a large-scale multimodal medical database," *BioData Mining*, vol. 17, no. 1, July 2024, Art. no. 22, https://doi.org/10.1186/s13040-024-00373-1.

[7] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "Prescriptive Method for Optimizing Cost of Data Collection and Annotation in Machine Learning of Clinical Ultrasound," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia, July 2023, pp. 1–4, https://doi.org/10.1109/EMBC40787.2023.10340858.

[8] Y. Nomura *et al.*, "Performance changes due to differences among annotating radiologists for training data in computerized lesion detection," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, no. 8, pp. 1527–1536, Aug. 2024, https://doi.org/10.1007/s11548-024-03136-9.

[9] F. Galbusera and A. Cina, "Image annotation and curation in radiology: an overview for machine learning practitioners," *European Radiology Experimental*, vol. 8, no. 1, Feb. 2024, Art. no. 11, https://doi.org/10.1186/s41747-023-00408-y.

[10] K. Zhang *et al.*, "Rep-GLS: Report-Guided Generalized Label Smoothing for Robust Disease Detection." arXiv, 2025, https://doi.org/10.48550/ARXIV.2508.02495.

[11] "RadLex radiology lexicon," *RAdiological Society of North America*. https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon.

[12] "SNOMED International - Service Migration." https://static-web.snomedtools.org/html/migrate.html.

[13] J. J. Budovec, C. A. Lam, and C. E. Kahn, "Informatics in Radiology: Radiology Gamuts Ontology: Differential Diagnosis for the Semantic Web," *RadioGraphics*, vol. 34, no. 1, pp. 254–264, Jan. 2014, https://doi.org/10.1148/rg.341135036.

[14] R. Chen, Z. Zhao, M. Yusufu, X. Shang, D. Shi, and M. He, "Choroidal Vessel Segmentation on Indocyanine Green Angiography Images via Human-in-the-Loop Labeling." arXiv, 2024, https://doi.org/10.48550/ARXIV.2406.01993.

[15] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "A human-in-the-loop method for pulmonary nodule detection in CT scans," *Visual Intelligence*, vol. 2, no. 1, July 2024, Art. no. 19, https://doi.org/10.1007/s44267-024-00052-z.

[16] M. Tounsi, E. Aram, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "A Comprehensive Review on Biomedical Image Classification using Deep Learning Models," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19538–19545, Feb. 2025, https://doi.org/10.48084/etasr.8728.

[17] A. Amalki, K. Tatane, and A. Bouzit, "Deep Learning-Driven Ontology Learning: A Systematic Mapping Study," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20085–20094, Feb. 2025, https://doi.org/10.48084/etasr.9431.

[18] F. M. Aldhafeeri, "Governing Artificial Intelligence in Radiology: A Systematic Review of Ethical, Legal, and Regulatory Frameworks," *Diagnostics*, vol. 15, no. 18, Sept. 2025, Art. no. 2300, https://doi.org/10.3390/diagnostics15182300.

[19] J. H. Lee, H. Hong, G. Nam, E. J. Hwang, and C. M. Park, "Effect of Human-AI Interaction on Detection of Malignant Lung Nodules on Chest Radiographs," *Radiology*, vol. 307, no. 5, June 2023, Art. no. e222976, https://doi.org/10.1148/radiol.222976.

[20] B. N. Patel *et al.*, "Human–machine partnership with artificial intelligence for chest radiograph diagnosis," *npj Digital Medicine*, vol. 2, no. 1, Nov. 2019, Art. no. 111, https://doi.org/10.1038/s41746-019-0189-7.

[21] A. Patil *et al.*, "Efficient quality control of whole slide pathology images with human-in-the-loop training," *Journal of Pathology Informatics*, vol. 14, 2023, Art. no. 100306, https://doi.org/10.1016/j.jpi.2023.100306.

[22] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning in medical image analysis," *Medical Image Analysis*, vol. 95, July 2024, Art. no. 103201, https://doi.org/10.1016/j.media.2024.103201.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 157–173, May 2008, https://doi.org/10.1007/s11263-007-0090-8.

[24] N. Fiedler, M. Bestmann, and N. Hendrich, "ImageTagger: An Open Source Online Platform for Collaborative Image Labeling," in *RoboCup 2018: Robot World Cup XXII*, vol. 11374, D. Holz, K. Genter, M. Saad, and O. Von Stryk, Eds. Springer International Publishing, 2019, pp. 162–169.

[25] "DICOM Web Viewer." https://ivmartel.github.io/dwv/.

[26] E. Ziegler *et al.*, "Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research," *JCO Clinical Cancer Informatics*, vol. 4, Apr. 2020, https://doi.org/10.1200/CCI.19.00131.

[27] "A Data Platform for Medical AI that enables building high-quality datasets and algorithms with lean process | VinDr," Mar. 26, 2021. https://vindr.ai/vindr-lab.

[28] A. Diaz-Pinto *et al.*, "MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images," *Medical Image Analysis*, vol. 95, July 2024, Art. no. 103207, https://doi.org/10.1016/j.media.2024.103207.

[29] C. Marzahl *et al.*, "EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control," *Scientific Reports*, vol. 11, no. 1, Feb. 2021, Art. no. 4343, https://doi.org/10.1038/s41598-021-83827-4.

[30] U. Rubens *et al.*, "Cytomine: Toward an Open and Collaborative Software Platform for Digital Pathology Bridged to Molecular Investigations," *PROTEOMICS – Clinical Applications*, vol. 13, no. 1, Jan. 2019, Art. no. 1800057, https://doi.org/10.1002/prca.201800057.

[31] R. Miao, R. Toth, Y. Zhou, A. Madabhushi, and A. Janowczyk, "Quick Annotator: an open-source digital pathology based rapid image annotation tool," *The Journal of Pathology: Clinical Research*, vol. 7, no. 6, pp. 542–547, Nov. 2021, https://doi.org/10.1002/cjp2.229.

[32] "Leading Image & Video Data Annotation Platform" *CVAT,* https://www.cvat.ai.

[33] "Open Source Data Labeling," *Label Studio*. https://labelstud.io/.

[34] J. Jäger, G. Reus, J. Denzler, V. Wolff, and K. Fricke-Neuderth, "LOST: A flexible framework for semi-automatic image annotation." arXiv, 2019, https://doi.org/10.48550/ARXIV.1910.07486.

[35] "imglab - Image processing and optimization on the fly," *imglab*. https://imglab.io/.

[36] "MedTAG - Medical Annotation Tool for Diagnostic Reports," *GitHub*. https://github.com/MedTAG.

[37] D. Strick, C. Garcia, and A. Huang, "Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification." arXiv, 2025, https://doi.org/10.48550/ARXIV.2505.06646.

[38] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017, pp. 3462–3471, https://doi.org/10.1109/CVPR.2017.369.

[39] "Radiology Lexicon | NCBO BioPortal." https://bioportal.bioontology.org/ontologies/RADLEX.

[40] "Human Phenotype Ontology | NCBO BioPortal." https://bioportal.bioontology.org/ontologies/HP.

[41] "CXR8 | Powered by Box." https://nihcc.app.box.com/v/ChestXray-NIHCC.

[42] "Random Sample of NIH Chest X-ray Dataset." https://www.kaggle.com/datasets/nih-chest-xrays/sample.