# MEDebiaser: A Human-AI Feedback System for Mitigating Bias in Multi-label Medical Image Classification

Shaohan Shi
ShanghaiTech University
Shanghai, China
shishh2023@shanghaitech.edu.cn

Yuheng Shao
ShanghaiTech University
Shanghai, China
shaoyh2024@shanghaitech.edu.cn

Haoran Jiang
ShanghaiTech University
Shanghai, China
jianghr2023@shanghaitech.edu.cn

Yunjie Yao
ShanghaiTech University
Shanghai, China
yaoyj2024@shanghaitech.edu.cn

Zhijun Zhang
Shuguang Hospital Affiliated to Shanghai University of Chinese Traditional Medicine
Shanghai, China
zjzhang2007@sina.com

Xu Ding*
Shuguang Hospital Affiliated to Shanghai University of Chinese Traditional Medicine
Shanghai, China
xu.ding2018@outlook.com

Quan Li*
ShanghaiTech University
Shanghai, China
liquan@shanghaitech.edu.cn

## Abstract

Medical images often contain multiple labels with imbalanced distributions and co-occurrence, leading to bias in multi-label medical image classification. Close collaboration between medical professionals and machine learning practitioners has significantly advanced medical image analysis. However, traditional collaboration modes struggle to facilitate effective feedback between physicians and AI models, as integrating medical expertise into the training process via engineers can be time-consuming and labor-intensive. To bridge this gap, we introduce *MEDebiaser*, an interactive system enabling physicians to directly refine AI models using local explanations. By combining prediction with attention loss functions and employing a customized ranking strategy to alleviate scalability, *MEDebiaser* allows physicians to mitigate biases without technical expertise, reducing reliance on engineers, and thus enhancing more direct human-AI feedback. Our mechanism and user studies demonstrate that it effectively reduces biases, improves usability, and enhances collaboration efficiency, providing a practical solution for integrating medical expertise into AI-driven healthcare.

## CCS Concepts

• **Human-centered computing** → Human computer interaction (HCI).

*Quan Li and Xu Ding are the corresponding authors.


*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

## Keywords

Multi-label Classification, Medical Image, Machine Learning, Interactive Systems and Tools

**ACM Reference Format:**
Shaohan Shi, Yuheng Shao, Haoran Jiang, Yunjie Yao, Zhijun Zhang, Xu Ding, and Quan Li. 2018. MEDebiaser: A Human-AI Feedback System for Mitigating Bias in Multi-label Medical Image Classification. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 27 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Artificial Intelligence (AI) has significantly enhanced healthcare providers' workflows in various areas such as electronic medical records management [57, 76], clinical decision support [4, 80], and particularly medical image analysis [6, 9, 28, 90]. AI systems aid in the analysis of medical images, including X-rays, CT scans, and MRIs, to detect lesions and abnormalities, thereby increasing diagnostic accuracy and efficiency [45, 56, 98].

When developing AI systems for medical image analysis, machine learning (ML) practitioners (hereafter "engineers") rely on medical professionals (hereafter "physicians") to provide a substantial number of accurately labeled or even annotated images for training purposes. These images need to precisely represent various diseases or conditions, allowing AI models to effectively differentiate between different medical scenarios. In many medical contexts, a single image may exhibit features of multiple conditions, a phenomenon known as *Multi-label Image* [21, 46, 100]. When building a system to handle multi-label medical images, engineers face two primary challenges inherent in medical datasets. The first challenge is *Imbalanced Distribution* [89]. Common diseases are overrepresented in the data, while rare diseases have fewer cases, leading to an imbalanced distribution of labels. The second challenge is *Label Co-occurrence* [12]. Certain symptoms often appear

together, creating complex dependencies between labels. For instance, in chest X-rays, conditions like cardiomegaly, consolidation, and edema often occur simultaneously [77], which can lead the model to confuse these co-occurring symptoms as variations of the same issue. These challenges introduce biases during training, making it difficult for the model to accurately differentiate between symptoms. This, in turn, affects the accuracy of diagnostic results and can undermine physicians' trust in AI systems.

The conventional approach to building MLMIC models involves a three-stage process: *data preparation*, *model training*, and *result verification* [58] (Figure 1). Physicians annotate images, engineers build and train models, and then physicians provide feedback on suboptimal results. This "Bias⇆Revision" cycle can be inefficient due to different work styles and perspectives, resulting in communication gaps and increased complexity in model iteration [56]. While more advanced interactive and visualization tools have been developed to involve physicians more directly, they have their own drawbacks. Engineers often utilize visualization and interactive techniques to facilitate such feedback [1, 93, 102], involving physicians in *data preparation* [5, 51], *model training* [65], and *result verification* [99]. These approaches often include well-designed interactive features, but require physicians to grasp fundamental AI and ML concepts, creating a steep learning curve for those without a computer science background. Moreover, while physicians may find the model's explanations satisfactory, this does not necessarily ensure that the model remains practical as human knowledge continues to be integrated. Thus, joint evaluation by physicians and engineers is essential: physicians should review the model's interpretability for accuracy, while engineers analyze the data to ensure the model's ongoing usability.

The rapid integration of AI into medical image analysis has also fostered interdisciplinary collaboration between physicians and engineers [4, 96]. Physicians contribute valuable data and provide critical insights, while engineers leverage advanced AI technologies to apply this knowledge. This synergy not only amplifies the application of AI in the medical field but also promotes its widespread adoption, thereby revolutionizing the way medical diagnostics and treatments are approached [56]. However, significant hurdles impede effective collaboration in MLMIC. During *data preparation*, engineers often rely on physicians to annotate various labels for each image in the dataset, i.e., specifying the features the ML system should learn to recognize. This approach, while essential, can be labor-intensive, especially when dealing with large datasets. In *model training*, state-of-the-art models have shown remarkable performance [40, 43, 47, 48, 60, 66, 73, 83, 91], yet their direct application in clinical settings remains challenging due to physicians' limited ML expertise. Physicians typically engage passively, just relying on pre-developed solutions created by engineers [59]. For *result verification*, when model performance is suboptimal—such as struggles with symptom differentiation or misidentification—physicians provide feedback for engineers, helping to refine the model. Due to the scarcity of medical data, expanding the dataset is often difficult, leading engineers to use mathematical or algorithmic approaches to align physicians' expertise with model predictions. However, translating medical expertise into mathematical formulations is complex.

To address these issues, our work focuses on two interconnected challenges in MLMIC: technical hurdles like imbalanced distributions and label co-occurrences, and the inefficient "Bias⇆Revision" collaboration loops that result from workflow and perspective mismatches between physicians and engineers. We introduce *MEDebiaser* (Figure 1), a human-in-the-loop system designed to bridge this gap by streamlining the revision cycle. This approach draws on the principles of mixed-initiative user interface design [41], aiming to achieve a seamless balance between automated assistance and human expertise. Unlike previous approaches, our method clearly defines the roles and responsibilities of physicians and engineers, aligning with the principle of considering uncertainty about user goals ([41], Principle 2). Instead of requiring physicians to annotate all images before training, the model first learns independently, and physicians can then use a user-friendly interface in *MEDebiaser* to focus specifically on annotating biased images, thereby reducing the workload associated with full data annotation, and minimizing the costs associated with incorrect predictions ([41], Principle 8), which replaces the inefficient back-and-forth with a shared interface that enables direct physician annotations and real-time model updates. This approach also empowers physicians to move beyond a passive feedback role; they can actively use familiar annotation methods to correct the model's understanding of biased images, directly incorporating their clinical expertise. Simultaneously, engineers are repositioned from constant, manual tuning into a supervisory role. While physicians interact with a simplified interface focused on accuracy and visual feedback, engineers can utilize a dedicated metrics dashboard for comprehensive evaluation and validation of the physician-driven updates. The system integrates their specialized knowledge only when relevant, enabling efficient agent-user collaboration ([41], Principle 10), thus avoiding unnecessary involvement in the complex modeling process. This clear division of roles ensures that clinical expertise is effectively integrated while model reliability is maintained, significantly enhancing cross-disciplinary collaboration without overburdening either party. This targeted involvement not only effectively mitigates bias in MLMIC but also reduces professional barriers and minimizes the need for frequent communication, streamlining physician-engineer collaboration. As a result, both the physical and cognitive workload for physicians and engineers are significantly reduced. In summary, the key contributions of this study are as follows:

- Through a *formative study*, we analyzed and summarized the collaborative modeling patterns and challenges in MLMIC from the perspectives of both physicians and engineers.
- We introduced *MEDebiaser*, an iterative system that redefines physician-engineer collaboration for bias mitigation in MLMIC by providing physicians with an interactive interface to directly address complex technical biases, thereby moving beyond inefficient revision cycles and better integrating expert medical knowledge.
- A *mechanism study* and a *user study* demonstrated that *MEDebiaser* significantly reduces bias and optimizes physician-engineer collaboration, confirming its effectiveness and applicability in the medical field.

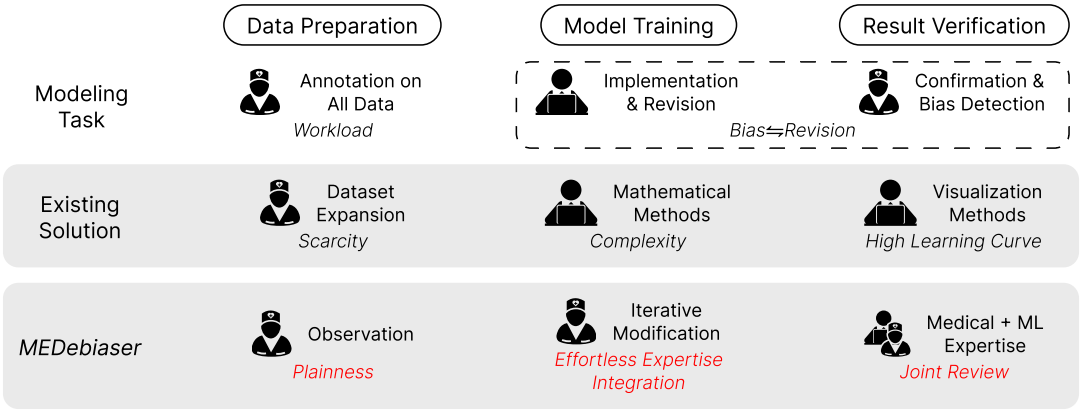


**Figure 1: Traditional modeling tasks face challenges such as the high workload of annotating entire datasets and the inefficiencies of the "Bias⇆Revision" mode. Existing solutions focus on dataset expansion—difficult to apply in the case of imbalanced data—or complex visualization and interaction systems. Additionally, integrating knowledge through engineers can be cumbersome and inefficient. *MEDebiaser* overcomes these by providing an accessible visual interface that allows physicians to apply their expertise in familiar ways, ensuring continuous monitoring by both engineers and physicians and fostering seamless collaboration between the two.**

## 2 Related Work

### 2.1 Multi-label Classification and Biases

Multi-label classification deals with instances that possess multiple labels simultaneously [32, 52]. In real-world scenarios, natural image data often exhibits a form of bias [81, 82]. This bias typically stems from low-quality training data, especially noisy labels [97]. Therefore, many approaches focus on addressing this issue at the dataset level [92, 101]. For instance, Reweighter [92] is a visual analytics tool that mitigates label quality issues through sample reweighting. During model training, bias primarily arises from an *Imbalanced Distribution* [3, 18], where the training set is highly skewed towards particular labels. Traditional resampling methods are mostly based on a single-label setup [53, 87]. Additionally, Ridnik et al. [70] adjusted the focal loss functions to create a dynamically balanced training process. Despite these techniques mitigating the problem of imbalanced distribution, AI models still face the issue of *Label Co-occurrence*, where images in the training set frequently feature contextual objects that co-occur with particular labels [37]. For example, images containing balls might be incorrectly labeled as dogs because balls and dogs often appear together. This issue is defined as contextual bias in some natural image datasets [27]. Several studies employ mathematical or algorithmic approaches to mitigate contextual bias. With the development of ML, Graph convolutional networks have been widely applied to leverage graph structure information, learning label features within the graph to enhance feature representation, thereby improving classification accuracy and effectiveness [15–17, 94].

In medical image classification, the multi-label bias problem has also garnered considerable attention [11, 12, 50, 104]. Unlike natural image datasets, the bias in medical images usually arises less from noisy labels, as they are mostly derived from diagnoses and reports by physicians [67]. Moreover, the scarcity of medical data makes addressing issues from a dataset perspective more challenging. During training, medical images exhibit similar phenomena to natural image datasets. Due to varying incidence rates of diseases, images of some rare symptoms are difficult to obtain [104]. Consequently, the *Imbalanced Distribution* in medical image datasets is more pronounced and severe, leading to poorer learning outcomes and higher error rates in identifying rare symptoms. In medical imaging, *Label Co-occurrence* becomes more complex than in natural image datasets. The co-occurrence of labels in medical datasets may arise from symptoms that often appear together, such as certain complications [84]. These symptoms may have very similar features in the images and be spatially close, making it difficult for the model to distinguish between different symptoms. The CXR-LT competition of the ICCV CVAMD 2023 Shared Task [39] aims to address the issue of long-tail distribution in multi-label thoracic disease classification in chest X-rays, with solutions delivering competitive results that somewhat alleviate the bias in medical multi-label images. In this challenge, Kim et al. [48] proposed a solution called *CheXFusion*, a transformer-based fusion model, which improves classification accuracy by effectively integrating features extracted from multi-view medical images using self-attention and cross-attention mechanisms.

Our work leverages pre-trained models for multi-label classification, specifically addressing challenges like *Imbalanced Distribution* and *Label Co-occurrence* in medical image datasets through a human-AI interactive feedback mechanism. Our method can be applied to various attention-based algorithms, enhancing their performance by improving the model's ability to detect complex patterns, increasing interpretability, boosting adaptability, and clinical applicability.

### 2.2 Human-AI Collaboration in Medicine

Human-AI collaboration (HAI) refers to the cooperative effort between humans and AI systems to solve specific problems, leveraging each party's strengths with clearly defined roles [55]. As HCI continues to evolve, there are currently three main modes of HAI [74]. The first mode is *AI-assisted Decision-making*, where AI acts as an

assistant providing recommendations to users, who then combine these suggestions with their prior knowledge to make the final decision. This mode is widely used in decision support systems across various fields, including finance [19], operations research [31], and healthcare [49, 57]. The second mode is *Human-in-the-loop*, where users intervene in the model training process to improve its performance. Interactive machine learning [23] exemplifies this approach. Yimam et al. [95] used interactive learning to automatically annotate medical documents. Guo et al. [30] proposed an interactive machine-learning approach for automatically grouping medical images, which was notably enhanced by incorporating expert-defined constraints. Moreover, Calisto et al. [8] found that tailoring an AI's communication style to the clinician's experience level in breast cancer diagnosis significantly reduced diagnostic time and errors. In *Joint Action*, users and AI work together as a team to achieve a shared goal, focusing on task allocation between the users and AI [55, 106, 107]. For example, *HADT* [105] uses a reinforcement learning-based strategy to decide human-machine task allocation in dialogue-based disease screening.

In the medical field, there is a growing emphasis on leveraging AI to enhance collaboration between physicians and technology across various tasks [79]. However, when using these tools, physicians frequently rely on engineers for troubleshooting and adjustments, as the modeling process can become time-consuming and challenging [59]. As a result, engineers play a crucial role in this human-AI collaboration, as their mathematical and algorithmic expertise is essential for refining AI models. Recently, in the fields of HCI and VIS, some work has allowed physicians to be directly involved in the modeling process. For example, Wang et al.[85] proposed *KMTLabeler*, a tool designed to involve physicians in labeling medical text. Similarly, Ouyang et al.[64] introduced a two-phase visualization system to facilitate interactive data analysis between physicians and AI. While these co-design approaches [22] effectively integrate medical expertise with AI, they often involve complex visualization and interactive interfaces and still require physicians to have a certain foundation of ML knowledge, which can lead to steep learning curves for physicians.

To address this gap, our system maintains the *Human-in-the-loop* mode, but it aims to redefine the role of AI from a mere tool to a bridge that enhances collaboration between physicians and engineers. We focus on designing straightforward, intuitive, and user-friendly interactions that, through interactive machine learning, enable physicians to engage more directly with the model. This approach not only reduces the workload of human parties but also optimizes the physician-engineer collaboration process.

## 2.3 Feedback in Interactive Machine Learning

Interactive machine learning is a paradigm where one or more users iteratively build and refine mathematical models through a cyclical process of input, review, and modification [44].

In interactive machine learning, the AI model processes inputs and generates appropriate outputs based on its understanding of the process that humans aim to represent. The feedback provided by AI to humans is facilitated through explainable AI (XAI). XAI has been applied across various fields to elucidate the inner workings of models, thereby increasing human trust in AI [1, 10, 59]. In image recognition tasks, local explanation is widely used in XAI due to its visual intuitiveness [24, 63, 75]. In medical imaging, Grad-CAM [72] is employed to visualize the productivity of AI models. Ouyang et al. [63] utilized Grad-CAM to generate saliency maps on MRI images, highlighting important areas due to its effectiveness in dealing with the distinct characteristics of medical images. In addition to XAI, visual analytics systems are frequently used to deliver AI-analyzed information to humans, exemplified by tools like *OoDAnalyzer* [13] and *VATLD* [29]. These systems leverage sophisticated interactive visualizations and advanced computational techniques to provide actionable insights, but are primarily designed for users with a background in machine learning or data science, which limits their accessibility to physicians without technical expertise.

Humans play a crucial role by contributing their domain expertise related to the data or models in interactive machine learning, thereby guiding model training. The feedback provided by humans, known as human input, is becoming increasingly integrated into AI research as part of human-AI collaboration. This integration primarily occurs through two modes: *Rule-based* and *Attention-guided* [27]. The *Rule-based* approach involves embedding predefined expert rules into the model before training to guide its learning. In medical image analysis, experts annotate prior knowledge beforehand. However, these expert rules often lack generality, and purely *Rule-based* methods may be insufficient for handling complex and dynamic scenarios [27]. Additionally, creating these rules requires significant human effort, such as annotating medical datasets, making this approach less practical and user-friendly for ML-naive users [23]. *Attention-guided* approaches, on the other hand, offer an effective strategy when embedding principles or rules is not feasible. The attention branch network [25] allows users to directly modify the model's attention on images, facilitating an interactive machine-learning process. This principle of guiding a model is foundational to the fine-tuning of modern architectures [2]. *GRADIA* [27] utilizes the interactive attention alignment framework to balance prediction accuracy and attention accuracy through human adjustments to model attention.

Current work on MLMIC has not fully leveraged the feedback between humans and AI to mitigate biases arising during model training. Our approach addresses this gap through an interface *MEDebiaser*. Existing attention-guided systems often require physicians to understand and manipulate an abstract representation of the model's focus, which can be unintuitive [14, 38]. In contrast, by using local explanations to demystify the model training process and present to physicians, *MEDebiaser* enables them to identify and correct biases directly before resuming training. Unlike traditional systems that often rely on complex and visually striking interactive designs, *MEDebiaser* is built on the principle of using methods that are familiar and intuitive to physicians. This approach prioritizes accessibility and usability, ensuring that the system is both effective and user-friendly for physicians.

# 3 Formative Study

The objective of our formative study is to comprehensively explore the perspectives of both physicians and engineers on *Multi-label Medical Image Classification* and *Human-AI Collaboration* tasks,

specifically addressing the following research questions: **RQ1:** *What are the current collaboration practices between physicians and engineers in building medical models, and what are the opportunities and challenges?* **RQ2:** *How do physicians and engineers perceive biases in multi-label medical images, and how have such biases been addressed previously together?*

To achieve this, we conducted *Semi-structured Interview* to delve into their challenges and expectations with institutional IRB approval. The insights gained from the interview enabled us to identify six key design challenges (**C1**-**C6**) and establish seven design goals (**DG1**-**DG7**).

## 3.1 Semi-structured Interview

*3.1.1 Participant.* We recruited five physicians (**D1**-**D5**) and four engineers (**P1**-**P4**) from local universities and hospitals (mean age = 34.11, SD = 5.23; 5 males and 4 females). Detailed participant information is provided in Table 1. All participants possess substantial experience in their respective fields and have collaborated on medical-related machine learning tasks with either physicians or engineers.

*3.1.2 Method.* Before the interview began, each participant signed an informed consent form that addressed privacy, ethics, and data collection for academic purposes. The semi-structured interview was conducted in a focus group format, with 9 participants engaging in the discussion for about 60 minutes. The sessions were moderated by an author, ensuring that each participant had an opportunity to share their perspectives.

The focus group discussion centered on the diverse viewpoints of physicians and engineers regarding MLMIC, including the unique characteristics and technical challenges of medical images. Additionally, the interview explored the collaboration modes between physicians and engineers, identifying potential challenges in their interaction. The discussion was audio-recorded with the participant's consent to facilitate accurate analysis later.

*3.1.3 Analysis.* We transcribed the audio recordings into text scripts and corrected transcription errors. Using thematic analysis [61], we conducted a comprehensive examination of the interview data. Initially, all authors read through the scripts to achieve a shared understanding, then proceeded with the coding process. During the initial coding phase, two authors segmented the text data into meaningful units and assigned labels to each segment. Subsequently, two other authors refined and adjusted these labels, uncovering additional patterns and themes. Finally, all authors shared and discussed the coding results, reaching a consensus to ensure consistency and reliability. Based on the coding results, we summarized the codebook (Table 2), which includes 4 themes and 13 labels, with definitions clarified based on the interviewees' perspectives. Through this rigorous process, we addressed **RQ1** and **RQ2**, summarizing the challenges and needs faced by physicians and engineers in MLMIC during their collaboration with AI.

## 3.2 Findings

*3.2.1 Challenges in Multi-label Medical Image Classification. Physician's Perspective.* All physicians underscored the unique complexities of multi-label medical images (N = 5). As **D2** pointed out, "*Medical images often have a lot of features of different lesions. Even for us, it takes a high level of expertise and careful attention, so I guess it's a real challenge for AI.*" Additionally, many physicians highlighted the issue of imbalanced label distribution in medical images (N = 3). According to **D1**, "*...a lot of symptoms are rare, and we might only see a few cases over several months, making it hard to capture them during [data] collection.*" Furthermore, several physicians discussed the phenomenon of *Label Co-occurrence* (N = 3). **D4** commented, "*Different lesions can be connected or show up together, like certain complications, especially in tongue images, which we might miss if we don't examine [them] carefully.*" These insights underscore the significance of *Imbalanced Distribution* and *Label Co-occurrence* as prevalent and critical issues in medical imaging. This leads to the first challenge: **C1. The intricate and multifaceted nature of multi-label medical images.**

*Engineer's Perspective.* Engineers believe that multi-label medical images face more severe challenges related to *Imbalanced Distribution* and *Label Co-occurrence* compared to multi-label natural images. They note that existing models and methods designed for multi-label classification may not perform well on medical datasets. **P4** observed, "*The data from physicians usually has a really obvious long-tail [distribution]. Plus, a lot of the features in the images are not only very similar but also close together, and they might even overlap.*" **P1** added, "*For these medical images, we usually hope that physicians can annotate them, but they're often not willing to spend a lot of time and effort on it.*" **P4** further commented, "*...if we could analyze the images at the pixel level, the model might pick up on the features better.*" This leads to the second challenge: **C2. Insufficient capability for fine-grained learning of the images.**

*3.2.2 Challenges in Human-AI Collaboration.* Both physicians and engineers acknowledge the benefits of incorporating human knowledge into the model training process, yet they recognize significant challenges (N = 8). Physicians highlighted that differences in terminology and understanding create communication barriers with engineers, a concern frequently supported by previous studies [56], which ultimately impacts the model's performance. **D1** explained, "*When I talk with collaborators, I try to simplify the medical terms, but they still often don't understand or misinterpret what I'm saying, which means we don't always get the AI results we're aiming for.*" Engineers echoed this sentiment, with **P2** stating, "*...these terms are new to us, so it's harder to grasp. We often struggle to fully understand what the physicians mean, which ends up affecting the functionality or outcomes they're looking for.*" **P3** further noted, "*Sometimes, a lot of the medical concepts or rules that physicians suggest are tough to translate into mathematical models during the process.*" Consequently, the medical knowledge or rules that physicians aim to integrate into the model are not always effectively communicated or implemented by engineers. This highlights the third challenge: **C3. The ineffectiveness and difficulty of integrating human knowledge into model training.**

Several physicians have reported a decrease in their trust in AI during collaboration (N = 3). This issue largely stems from the lack

**Table 1: The details of Semi-structured Interview participants.**

| ID | Gender/Age | Research Area | Experience | Title | Familiarity with AI |
|---|---|---|---|---|---|
| **D1** | Male/42 | Cardiothoracic Surgery | 17 Years | Clinical Professor | Aware |
| **D2** | Female/36 | Cardiothoracic Surgery | 12 Years | M.D. | Neutral |
| **D3** | Male/33 | Ear, Nose, and Throat | 7 Years | M.D. | Aware |
| **D4** | Female/31 | Traditional Chinese Medicine | 6 Years | Postdoc | Unfamiliar |
| **D5** | Male/29 | Orthopedic Surgery | 5 Years | Postdoc | Neutral |
| **P1** | Female/41 | Machine Learning | 15 Years | Associate Professor | Expert |
| **P2** | Male/37 | Machine Learning | 12 Years | Assistant Professor | Expert |
| **P3** | Male/31 | Human-Computer Interaction | 7 Years | Postdoc | Familiar |
| **P4** | Female/27 | Data Scientist | 5 Years | Ph.D. | Expert |

**Table 2: Codebook for Semi-structured Interview, targeting at RQ1 and RQ2.**

| Theme | Label | Definition |
|---|---|---|
| Collaboration | Physician's role | Roles and responsibilities of physicians |
| | Engineer's role | Roles and responsibilities of engineers |
| | AI Model's role | Roles and responsibilities of AI models |
| | Collaboration Workflow | Mode, steps, timing, and other details of collaboration |
| | Collaboration Quality | Challenges, outcomes, feelings, and evaluations of collaboration |
| Multi-label | Data Quality | Accuracy & completeness of datasets and distribution & co-occurrence of labels |
| | Medical Features | Special characteristics of medical multi-label images |
| | Potential Biases | Potential biases that may arise from using medical multi-label images |
| AI Model | Model Performance | Accuracy, validity, and stability of model |
| | Model Explainability | Clear presentation of model results and decision processes |
| | Knowledge Integration | Opinions and concerns on integrating human knowledge into model training |
| Expectation | Interactivity | Interface's ability to facilitate Human-AI collaboration |
| | Feedback | Clear presentation of model metrics and improvements |

of model interpretability, a concern also highlighted in previous research [1]. As **D3** remarked, "*I don't get how the model comes to these conclusions, and sometimes I can't be sure its decisions match our clinical judgment.*" Additionally, mistrust can arise from physicians not being involved in the model training and decision-making processes. **D2** pointed out, "*Usually, I just get the results from the model, but if I spot [any] errors, I have no way to fix them. This really hurts my trust in the system.*" Consequently, the collaboration suffers from inadequate feedback between the parties, which significantly erodes the physicians' trust. The fourth challenge is: **C4. Lack of mutual feedback in the collaboration between physicians and AI models.**

When discussing the collaboration between physicians and engineers, both parties raised several specific and nuanced challenges. One major concern was the difference in their working styles (N = 6). **D4** emphasized, "*Our work is usually driven by clinical needs and urgency, while they're more focused on project deadlines and data availability. This difference often causes mismatched timelines and inefficiency.*" **P1** added, "*Physicians' time is incredibly valuable, and they want to see results quickly, but developing and training our models takes time and constant adjustments.*" In addition, **D5** mentioned, "*Previous collaborators gave us some interfaces to help with data analysis or using AI, but I found them hard to use, and some features were tough to understand even after training.*" The complexity of existing workflows often demands significant time and effort from physicians to learn and engage in the modeling process, underscoring the fifth challenge: **C5. Inconsistency and high learning costs in the collaboration between physicians and engineers.**

Furthermore, engineers pointed out that physicians might experience issues related to the "overuse" of AI models (N = 2). **P1** observed, "*In a previous project, there was a case where the model*

*picked up incorrect information because physicians used it improperly, so it's necessary for us to re-evaluate the model's usability.*" Similarly, **D5** mentioned, "*I'm concerned I might make mistakes without even realizing it, so I prefer to double-check the performance with engineers.*" This highlights the sixth challenge: **C6. Lack of joint review of the model's feedback by physicians and engineers.**

## 3.3 Design Goals

After interviewing both physicians and engineers, we have summarized a series of design goals aimed at effectively addressing the aforementioned challenges. Our approach focuses on improving the traditional collaboration mode between these two parties (Figure 1), enhancing conventional classification methods to foster more direct feedback and interaction between physicians and AI models. We have organized the seven design goals into two key areas: Bias Refinement, which involves *Pixel-level Bias Refinement Integrated with Physicians' Knowledge*, and Bias Explanation, which focuses on the *Display and Analysis of Biases During AI Training and Fine-tuning Results*. These components function within an Iterative Loop, continuously reducing bias, refining the model's predictive performance, increasing physicians' engagement and trust, and simultaneously lightening the workload of engineers.

Bias Refinement Given the differences in professional expertise, work modes, and technical barriers between physicians and engineers, efforts to incorporate medical knowledge into AI training may not always produce optimal results, leading to inefficiencies in collaboration. However, this integration is essential for effective medical image training. To overcome these challenges (**C3**, **C5**), we aim to enhance more direct cooperation between physicians and AI models, while also improving physician-engineer collaboration. This brings us to our first design goal: **DG1. Facilitate more direct interaction and knowledge integration between physicians and AI models.** This objective was emphasized repeatedly during our semi-structured interview.

Bias Explanation To address the fourth challenge (**C4**), we should not only facilitate the integration of physicians' knowledge into AI but also provide meaningful feedback from the AI back to the physicians. This requires clear displays of the model's training and decision-making processes, enabling physicians to identify biases or inaccuracies generated by the model. For multi-label medical images, it is essential to present the different symptoms within a single image. As **D3** emphasized during the semi-structured interview, "*I want the model to clearly mark or highlight the locations of each symptom it identifies. I will only trust it when its judgments align with mine.*" This underscores the importance of our second design goal: **DG2. Provide detailed displays of the model's decision-making process during training.**

Bias Explanation Beyond providing feedback during AI training, physicians also emphasized the importance of displaying the outcomes of the training process. As **D2** mentioned, "*I want to see what changes occur after I correct errors and whether these changes improve the model's performance.*" Illustrating how physicians' interventions affect model accuracy and reduce bias allows them to better understand the impact of their contributions, which in turn helps validate the model's usability. This brings us to our third design goal: **DG3. Display the impact of physicians' interventions on model outcomes.** This goal further supports the feedback mechanisms outlined in **C4** and **C6**.

Bias Explanation While physicians play a crucial role in verifying the model, the expertise of engineers in evaluating and controlling the model is equally essential. Even when physicians find the model's interpretability satisfactory, risks such as overfitting[1] or gradient vanishing[2] could still arise due to excessive bias refinement or inherent model issues. Identifying and addressing these potential problems to keep the model usable requires engineers' specialized knowledge. Therefore, it is vital for both physicians and engineers to collaboratively validate the model. To address **C6** and supplement **DG3**, our fourth design goal is: **DG4. Provide feedback on model performance to engineers.**

Bias Refinement To mitigate issues of *Imbalanced Distribution* and *Label Co-occurrence* in MLMIC (**C1**), engineers typically rely on mathematical and algorithmic techniques. As **P2** noted, "*The model's inaccuracies often stem from focusing on incorrect areas.*" By enabling physicians to correct biases that arise during model training before optimization is passed on to engineers, the model's performance and effectiveness can be significantly improved. However, for ML-naive users, steering AI models can be challenging, as highlighted in **C5**. Therefore, it's essential to first provide physicians with an understanding of the data and then offer intuitive, user-friendly tools that enable them to easily identify and correct biases. Our fifth design goal is: **DG5. Provide physicians with data cues and accessible tools to adjust and correct biases during model training.**

Bias Refinement The similarity of features and the spatial proximity or overlap in medical images pose significant challenges for effective training. Engineers have observed that for a model to accurately learn the characteristics of various symptoms, it requires more detailed image analysis (**C2**). However, the annotations needed for this level of detail depend heavily on the expertise of physicians, and such annotated data is often lacking. To address this, we can involve physicians in the annotation process during training, allowing them to highlight key features in the images where detailed learning is required. Thus, our sixth design goal is: **DG6. Incorporate techniques that enable the model to discern and learn subtle features in medical images.**

Iterative Loop Our final design goal is: **DG7. Develop an interactive interface that facilitates iterative feedback between physicians and AI models.** This goal is crafted to support feedback exchange (**C4**) and foster more direct and closer interactions (**C3**) between physicians and AI models. Within this iterative loop, the process of physicians providing feedback to the AI and receiving feedback from the AI is seamlessly integrated. Through continuous interaction, physicians can incrementally reduce model biases and enhance their performance. The importance of this goal lies in its potential to gradually increase physicians' familiarity with human-AI collaboration, allowing them to progressively refine biases and improve the model's performance.

[1] Overfitting happens when training data details are learned too specifically, so performance drops on new data.
[2] Gradient vanishing occurs when gradients become very small, slowing or stopping learning.

### 3.4 Content Analysis

To develop a user-friendly and accessible annotation method (**DG5**), we first conducted a comprehensive review of the annotation tools commonly employed by our target users. This review included mainstream annotation tools identified through an extensive search across online platforms such as YouTube and GitHub, recent peer-reviewed literature, and interviews with medical imaging experts. As a result, we identified 5 annotation methods frequently employed in medical AI annotation workflows. We then systematically analyzed their strengths and limitations (Table 9, Appendix A), evaluating them based on usability, annotation granularity, and their suitability for medical applications.

To empirically evaluate these methods, we designed and conducted a usability study in which participants annotated a sample dataset using each of the identified techniques. The results, summarized in Table 10, Appendix A, include both quantitative performance metrics and qualitative feedback. Based on the study's findings and expert consultations, we identified polygon annotation as the optimal method, striking a balance between annotation accuracy and usability for physicians.

## 4 MEDebiaser

Based on the identified user needs, we designed and implemented *MEDebiaser* (Figure 2), an interactive system that enhances collaboration between physicians and engineers in MLMIC, addressing the following research question: **RQ3:** *What practices do physicians and engineers hope to adopt to mitigate multi-label issues and enhance their collaboration?*

*MEDebiaser* features five main components: the *Panel View* (Figure 2-(A)) allows users to upload datasets, select models, and adjust training parameters (**DG1**); the *Label View* (Figure 2-(B)) displays the distribution of labels in a table and their co-occurrence (**DG5**); the *Attention View* (Figure 2-(C), Bias Explanation) provides local explanations during the training and fine-tuning process (**DG2**, **DG7**); the *Modification View* (Figure 2-(D), Bias Modification) offers intuitive tools for pixel-level fine-tuning (**DG1**, **DG5**, **DG6**, **DG7**); and the *Performance View* (Figure 2-(E), Bias Explanation) displays prediction results and the impacts of user interventions on model outcomes (**DG3**, **DG4**). Users can then initiate a new round of fine-tuning based on the retrained model, following an Iterative Loop process. The *MEDebiaser* workflow (Figure 3) is structured into three main stages: *Loading Dataset and Model*, *Observing and Modifying Attention*, and *Evaluating Model Performance*. By iteratively executing these stages, *MEDebiaser* enables physicians to continuously correct biases that arise during model training, seamlessly incorporating their medical expertise into the training process. To be noted, the actual application of the system is not restricted to the specific dataset and model showcased.

### 4.1 Loading Dataset and Model

At this stage, the users can upload datasets, review label distribution and co-occurrence within the dataset, select the model to be used, and configure the appropriate training parameters.

*4.1.1 Dataset.* To evaluate the system, we conducted tests on several datasets, with **ChestX-ray14** [88] being one of the most representative examples. This dataset includes 112, 120 frontal view chest X-ray images, each at a resolution of $1024 \times 1024$, collected from 30, 805 patients between 1992 and 2015. These images have been automatically labeled into 14 different thoracic pathology categories based on radiological diagnostic reports, with each image potentially containing multiple labels. As shown in Figure 9, Appendix B, the label distribution is highly imbalanced, with the most frequent label appearing nearly 20, 000 times and the least frequent only 227 times. Additionally, there is a noticeable co-occurrence among the labels, as shown in Table 11, Appendix B.

*4.1.2 Model.* To tackle the unique challenges of MLMIC, *MEDebiaser* employs a one-vs-rest classification strategy, treating the problem as N parallel binary classification tasks, where N is the number of possible labels. Architecturally, the system utilizes a powerful convolutional neural network (CNN) backbone—such as DenseNet [42] or ResNet [35]—which outputs an N-dimensional logit vector. Each logit is then passed through a sigmoid activation function, rather than the mutually exclusive softmax, to independently compute each class's probability. For training, the model minimizes the mean of N binary cross-entropy (BCE) losses against the multi-hot ground truth vector. This entire approach is inherently robust for handling common issues like *Imbalanced Distribution* and *Label Co-occurrence*. To ensure strong initial performance, the selected CNN models are pre-trained on extensive datasets like ImageNet [20] before training commences on the user's dataset.

*4.1.3 Interaction.* Initially, users can upload their datasets through the *Panel View* by clicking Upload (Figure 3-1). Once the dataset is uploaded, users can observe the proportion and distribution of each label in the *Label View* via a table (Figure 3-2). The sorting feature allows users to easily identify labels with smaller proportions, which may be more prone to biases and errors during model training and thus require special attention (**DG5**). Each label is color-coded in the table, and in the accompanying chord diagram (Figure 3-3), which is well-suited for showing data with associative relationships within the dataset [86]. Each arc connects to its co-occurring labels, with the band thickness representing the frequency of co-occurrence. This diagram enables users to clearly observe label co-occurrence and identify features that may need additional focus (**DG5**). Next, users can select a model from a predefined set using Select Model (Figure 3-4). To ensure the model can run on various machines, we provide commonly used CNN models for multi-label scenarios with default settings (batch size = 4, epoch = 30), which simplifies the technical setup and allows physicians to proceed with confidence without deep ML knowledge. These default parameters are highly versatile and compatible with different hardware environments, computational capabilities, and workflows in healthcare facilities. For physicians who need to modify parameters or are familiar with ML, clicking on ◂ allows direct adjustment of these parameters and models using  (Figure 3-5). This flexibility offers more diverse options to optimize training according to their specific needs, training time, and hardware requirements. The user then clicks Train (Figure 3-6), and

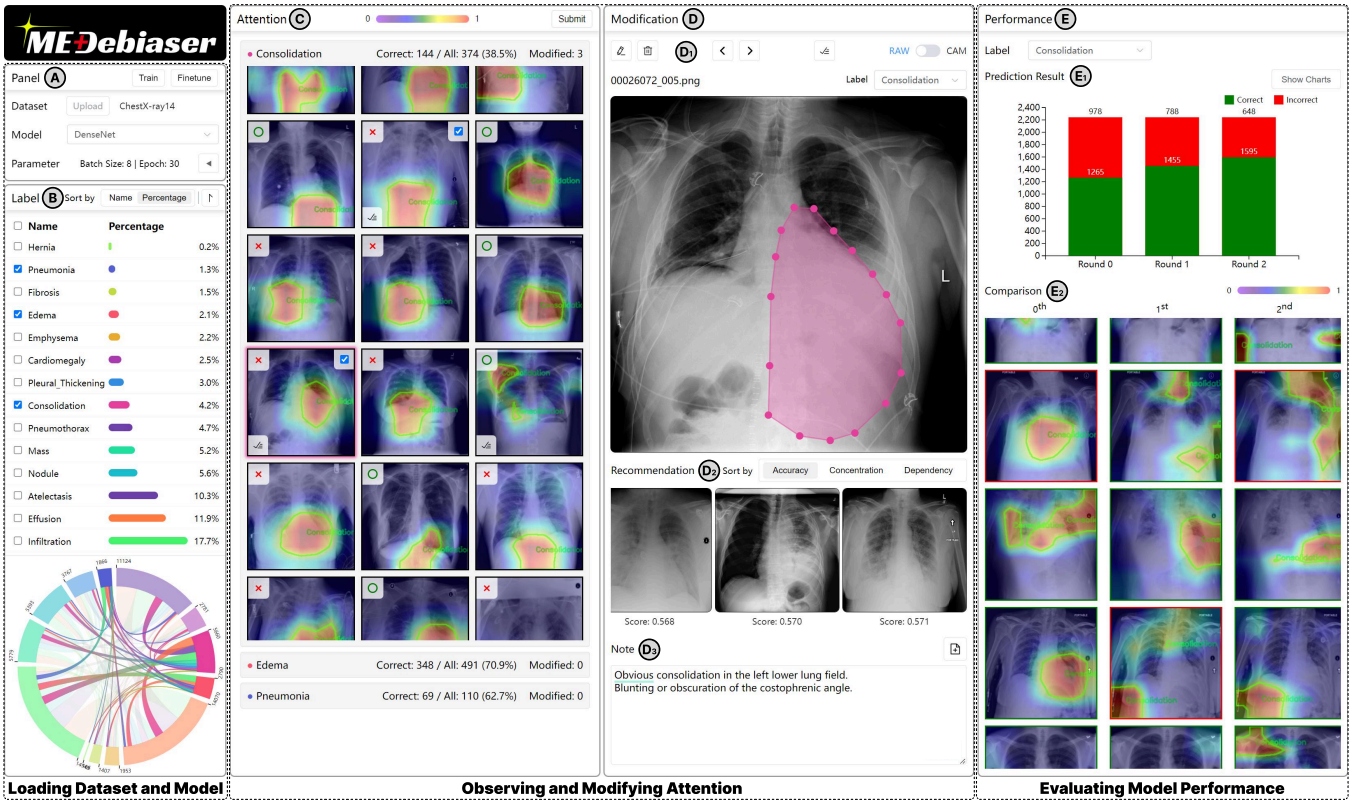


**Figure 2: (A) The *Panel View* provides components for uploading datasets, selecting models, and setting training parameters. (B) The *Label View* includes a table displaying label distribution and a chord diagram showing co-occurrence. (C) The *Attention View* displays local explanations for the selected labels. (D) The *Modification View* includes ($D_1$) an *Editing Area* for fine-tuning attention, ($D_2$) a *Recommendation Area* for sorting based on metrics, and ($D_3$) a *Note Area* for record-keeping. (E) The *Performance View* contains ($E_1$) a *Metrics Area* to show the model's metrics for each round of fine-tuning on specific labels, as well as ($E_2$) a *Comparison Area* to display comparisons of local explanations after each round of fine-tuning.**

the preliminary training begins. Upon completion, users can review the performance of each label in the *Performance View*. Labels that require further attention can be selected by clicking ☑ in the table or by clicking on the arcs in the chord diagram, which will then highlight them in the *Attention View*.

## 4.2 Observing and Modifying Attention

This stage is designed to achieve two main objectives: first, to present the model's decision-making process to the physicians (**DG2**); and second, to enable physicians to refine the model's attention on biased images by incorporating human expertise (**DG1**). This is accomplished through an iterative "observe-annotate-fine-tune" active learning loop. To address the first objective, the system provides *local explanation* by displaying Grad-CAMs on validation images. To streamline the review process, a *customized ranking strategy* then prioritizes the most problematic cases based on three active learning criteria: prediction accuracy, attention concentration, and co-occurrence dependency. Guided by this *ranking*, physicians can achieve the second objective by annotating correct regions using pixel-level masks. The model is subsequently *fine-tuned* with a dynamically weighted joint prediction. Importantly, Grad-CAM is leveraged here not just for interpretation, but for rapid error localization and correction, improving model performance with minimal expert effort while enhancing collaboration.

*4.2.1 Local Explanation.* During the preliminary training phase, models trained on the **ChestX-ray14** dataset are used to make predictions on the validation subset. Simultaneously, Grad-CAM is applied to the last convolutional layer of the CNN model, offering an effective XAI technique for interpreting CNNs [72]. Unlike traditional interpretability methods such as SHAP [54] and LIME [69], which can struggle with feature correlations or focus on small regions of the decision space, Grad-CAM leverages the natural structure of CNNs to highlight the most influential image regions [72]. This results in higher efficiency and better integration into rapid-feedback workflows. While Guided Backpropagation [78], a commonly used CNN visualization method, emphasizes pixel-level contributions, often amplifying noise in multi-label scenarios, Grad-CAM mitigates this by utilizing higher-level feature

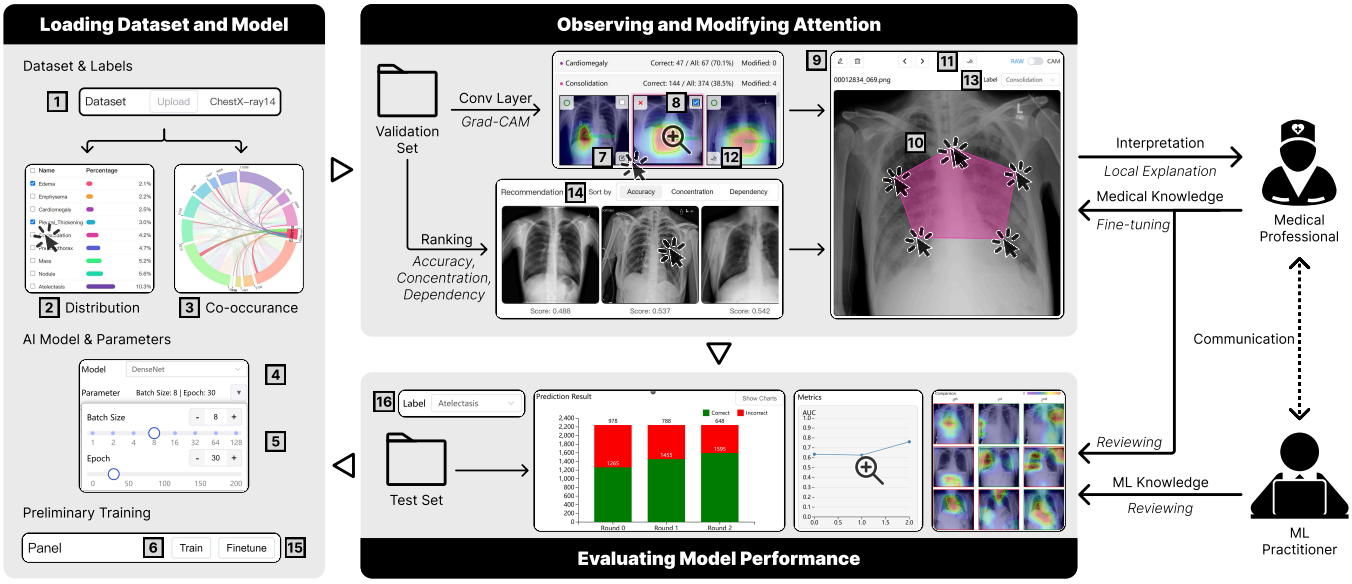


**Figure 3: The *MEDebiaser* workflow includes three main stages: *Loading Dataset and Model, Observing and Modifying Attention,* and *Evaluating Model Performance.***

maps, offering clearer and more intuitive explanations for physicians. Specifically, Grad-CAM calculates the gradients of a target label with respect to the feature maps of the selected convolutional layer. These gradients are then globally averaged to determine the importance weights for each feature map, highlighting the regions most influential in the model's predictions. The weighted combination of feature maps is visualized as a heatmap, overlaid on the original image to indicate the spatial areas crucial to the model's decision-making process. A comparison of the methods discussed above is provided in Figure 10, Appendix C.

*4.2.2 Fine-tuning.* In the modification phase of *MEDebiaser*'s workflow, physicians critically assess the model's heatmaps and provide pixel-level feedback by annotating incorrect attention areas with polygon masks. To incorporate this clinical feedback, we fine-tune the model using a dynamically weighted joint loss function. This function is composed of two distinct components: for the prediction loss, we adopt the function from the Explanation-guided Learning framework [26], an approach whose effectiveness is supported by studies like MAGI [103]. The attention loss, in contrast, is defined as the mean squared error between the model's Grad-CAM heatmap and the physician-provided mask, which helps the model focus on important features, improving accuracy. Crucially, the balance between these two losses is dynamically adjusted, with the weights for the attention loss being proportionally tailored based on the frequency of each label in the dataset. This combined approach ensures a balanced emphasis during training and effectively steers the model's focus toward the correctly annotated regions to optimize both prediction accuracy and explanation clarity.

*4.2.3 Ranking.* In the proposed iterative refinement process for *MEDebiaser*, active learning [68] is leveraged to enhance the efficiency of image annotation. The process begins by ranking images based on their likelihood of errors, with the model prioritizing those it finds most uncertain. This ranking strategy, a core principle of active learning, directs physicians' attention to the most critical cases—where the model's predictions are least certain or most prone to error. By focusing on these uncertain samples, physicians provide high-value annotations that significantly contribute to model improvement. This approach streamlines the identification and correction of potential inaccuracies early in the refinement cycle, enabling the model to learn from the most informative data and progressively enhance its accuracy with minimal human effort.

*Iterative Refinement Cycle.* The workflow follows a cyclical process where physicians start by annotating images predicted to be most susceptible to errors. The model is then fine-tuned with this new data, leading to the generation of updated heatmaps for the remaining images. These images are re-ranked based on the revised error likelihood assessments. This cycle continues, with each iteration progressively enhancing the model's accuracy and interpretability until the physicians determine that the model has reached a satisfactory level of performance.

*Customized Ranking Strategy.* To streamline the process, a three-tiered ranking system is implemented, tailored to align with the preferences and expertise of physicians:

- **Label Prediction Accuracy:** Experts suggest prioritizing images where the model's confidence significantly deviates from certainty. These images are ranked based on the deviation of their predictive values from the ideal, with values close to 1 indicating high accuracy and confidence, and values close to 0 representing low confidence and a higher likelihood of misclassification.
- **Heatmap Concentration:** The system analyzes the model's attention focus by examining the heatmap's $n \times n$ matrix, where $n$ represents the image's pixel dimensions. To ensure

a fair and comparable analysis across different images, each Grad-CAM heatmap is first normalized using global min-max scaling, which rescales all values to a range between 0 and 1. This normalization is critical as it allows for ranking based on relative attention deviation rather than absolute, unscaled gradient magnitudes. A value approaching 1 indicates a strong, concentrated focus on the most critical region, while a lower value suggests the model's attention is more diffuse.

- **Co-occurrence Matrix Dependency:** Experts believe that labels that frequently co-occur with other labels are more likely to be misclassified. To quantify this, the dependency between labels is evaluated using a co-occurrence matrix $M$, where $M_{ij}$ represents the frequency of co-occurrence between label $i$ and label $j$. To improve the discriminative power of this metric, we use inverse frequency instead of direct frequency, which reduces the dominance of ubiquitous labels and highlights more meaningful, strongly associated pairs. The inverse frequency for a target label $c$ with each label $j$ is computed as (1):

$$\text{inverse frequency}(c, j) = \frac{1}{M_{cj} + 0.01} \tag{1}$$

The small constant 0.01 ensures stability by preventing division by zero. The overall dependency score for the target label $c$ is then obtained by normalizing these inverse frequencies and summing over the relevant labels. To prevent rarely co-occurring labels from inflating the scores, we include only "positive labels"—those whose co-occurrence with $c$ exceeds a defined threshold—ensuring that weak or incidental associations have no impact on the final score. The dependency score is calculated as (2):

$$\text{dependency score}_c = \sum_{j \in \text{positive labels}} \frac{\text{inverse frequency}(c, j)}{\sum_k \text{inverse frequency}(c, k)} \tag{2}$$

A dependency score close to 1 indicates a high likelihood of misclassification due to strong interdependencies, while a score close to 0 suggests lower interdependency and thus fewer challenges in classification.

*4.2.4 Interaction.* For the first objective, the *Attention View* provides a clear display of the Grad-CAM visualization of the model's decision-making process on the validation set. Users can zoom in on any image by clicking on it, with icons on the top left indicating the accuracy of the model's label recognition— ◯ for correct predictions and × for incorrect ones. This allows physicians to assess whether the local explanations for the selected labels are reasonable, particularly for symptoms that are less frequent or often co-occur with others (**DG2**). If an explanation is deemed unreasonable, users can click ✎ (Figure 3-7) in the bottom right, which will bring the corresponding image into the *Modification View*. For those reasonably predicted, users can click ☑ (Figure 3-8) located in the top right corner. This action will save the Grad-CAM of the selected image as polygons, which will then be used in the next round of fine-tuning.

For the second objective, we employ annotation methods familiar to physicians [33], allowing them to mark areas of interest directly on the image using polygons, which is an intuitive process that does not require ML knowledge, thereby lowering the technical barrier for them to contribute their domain expertise (**DG5**). In the *Modification View*, users can enter annotation mode by clicking ✎ (Figure 3-9). They can then draw polygons by clicking on the image and double-clicking to finish the annotation (Figure 3-10). Annotating at the pixel level enables the model to focus on more detailed features (**DG6**). After completing the annotation, users can save their work by clicking ✓≡ (Figure 3-11), storing the polygon for the next round of fine-tuning. Annotated images will display ✓≡ (Figure 3-12) in the bottom left corner of the images in *Attention View*. Users can also differentiate between various labels within the same image by selecting from Select Label (Figure 3-13) and annotating different pixel-level areas accordingly.

In the *Recommendation Area* (Figure 2-D2), we prioritize images requiring attention by calculating label prediction accuracy, heatmap concentration, and co-occurrence matrix dependency. Users can choose one of these three methods to sort and focus on the images most in need of modification (Figure 3-14). Additionally, in the *Note Area*, users can take notes on the image. Compared to the traditional practice of annotating entire datasets, these designs significantly reduce the workload of physicians.

After labeling the biased images and saving the polygon information, users can click Finetune (Figure 3-15) to start a new round of fine-tuning (**DG7**). Once completed, the results will be updated in the *Performance View*.

## 4.3 Evaluating Model Performance

In this stage, users evaluate the model's effectiveness by analyzing parameters and Grad-CAM on the test set.

In the *Performance View*, users begin by selecting the label they wish to review from Select Label (Figure 3-16). The *Prediction Results* (Figure 2-E1) displays, for each round of fine-tuning, the number of correctly and incorrectly predicted labels on the test set. This visualization enables users to track the trend in correct predictions over time, providing insight into model performance progression. By observing the trends in these charts, physicians can assess whether their modifications are steering the model in a positive direction (**DG3**). Clicking on the Show Charts button reveals the evaluation metrics for each round, including *precision*, *recall* (or *sensitivity*), *F1 score*, and *AUC*. The rationale for selecting these specific metrics—while omitting others—is detailed in Table 12, Appendix D. To accommodate the varying levels of technical expertise among users, the system adopts a layered presentation strategy. While these parameters are not directly shown in the default view, they remain accessible through interaction for users with machine learning knowledge, such as engineers, to verify that the model remains functional and reliable (**DG4**). At the same time, physicians are provided with visual trends that reflect performance changes in a more intuitive and accessible manner, aligned with their diagnostic thinking patterns. This design intentionally abstracts away complex terminology, reducing cognitive load for non-expert users,

while preserving access to detailed metrics when needed. In doing so, the system strikes a balance between interpretability and transparency, fostering trust and effective collaboration between domain experts and technical practitioners.

In the *Comparison Area* (Figure 2-E2), test set images are displayed vertically, with the first column showing the Grad-CAM generated after the initial training, and each subsequent column presenting the Grad-CAM after each round of fine-tuning. Images with a green border indicate correct label predictions, while a red border indicates incorrect predictions. By observing these changes in local explanations, physicians can more intuitively understand how the focus of the model's attention evolves, gaining deeper insights into the model's decision-making process (**DG3**).

## 5 Evaluation

We evaluated *MEDebiaser* through two studies. First, the *mechanism study* validated that employing human-AI interactive feedback to mitigate existing biases can improve outcomes in the intended direction, providing a quantitative answer to the following research question: **RQ4:** *Does the system reduce bias in MLMIC and improve model performance, and how do physicians interact with the system to mitigate these biases?*

Second, after obtaining institutional IRB approval, we conducted a *user study*, which offered a comprehensive assessment of the user experience with *MEDebiaser*. Our primary goal was to determine whether *MEDebiaser* enhances MLMIC performance by mitigating the effects of *Imbalanced Distribution* and *Label Co-occurrence* and whether it can optimize physician-engineer collaboration, thereby addressing **RQ4** and the following research question: **RQ5:** *What's the impact of the system on workload, physician diagnosis, collaboration, and medical knowledge integration?*

### 5.1 Mechanism Study

In this section, we conducted two experiments to evaluate the effectiveness of *MEDebiaser* in mitigating model biases. The first experiment, *With and Without Attention*, compared models that incorporated attention modification on biased images with those that did not. The goal was to determine whether fine-tuning the model using a combination of prediction loss and attention loss effectively mitigated these biases. In the second experiment, *Breadth and Depth*, we explored four different annotation modes, varying in *depth* (few vs. many annotations) and *breadth* (random vs. focused labeling), to analyze how these strategies impact the model's generalization capabilities and its ability to reduce bias.

*5.1.1 Setup and Details.* Both experiments were conducted using the **ChestX-ray14** dataset and carried out on an NVIDIA 4070 Ti GPU. To reduce the annotation workload for physicians, we selected approximately one-tenth of the original dataset while preserving the label distribution and co-occurrence patterns. The data was split into an 8:1:1 ratio for training, validation, and testing. A DenseNet model pretrained on ImageNet was employed for both experiments, with fine-tuning performed using a learning rate of $1e-4$, a batch size of 4, and 30 training epochs. To maintain consistency and robustness, we applied standard data augmentation techniques across both experiments.

*5.1.2 Experiment I: With and Without Attention. Procedure.* We began by pretraining a DenseNet model on ImageNet, followed by initial training on the **ChestX-ray14** training set. The model generated chest X-ray images with Grad-CAM heatmaps and corresponding predictions for both the validation and test sets. Physicians were then tasked with reviewing these images, selecting 100 examples specifically focusing on the label Pleural_Thickening due to its rarity and higher error rate in prior predictions. Each selected image was meticulously annotated, with polygons drawn around regions of interest, covering all relevant labels.

Using these 100 annotated images, we employed two model fine-tuning techniques. The first method combined prediction loss with attention loss, as outlined in Section 4.2.2, leveraging Grad-CAM to refine the model's attention towards the correct regions. The second method relied solely on prediction loss, without any attention-based adjustments.

*Result & Analysis.* The results, as shown in Table 3, indicate that the model fine-tuned using the combined prediction loss and attention loss approach demonstrated notable improvements in handling the label Pleural_Thickening compared to using prediction loss alone. Specifically, the combined approach achieved a *precision* of 0.160, representing a 14.3% relative increase from 0.140 under prediction loss only. This improvement is also reflected in *recall* and *F1 score*. In terms of overall performance, *AUC* also showed a meaningful enhancement, improving from 0.613 with prediction loss only to 0.675 when incorporating attention loss. This suggests a stronger ability to distinguish between positive and negative cases. These quantitative gains underscore the impact of attention modification, which proved particularly beneficial for underrepresented, tail-end labels. By incorporating attention loss, the model was able to more effectively focus on the relevant image regions associated with such rare labels, reducing false negatives, and addressing the inherent challenges posed by imbalanced multi-label data.

*5.1.3 Experiment II: Breadth and Depth. Procedure.* We considered two dimensions for Experiment II: *breadth* and *depth*. *Breadth* refers to annotating a larger number of images, and *depth* focuses on consistently annotating images for a single label. The experimental setup closely mirrored that of Experiment I, utilizing four annotation strategies: 1) annotating 50 images with Pleural_Thickening; 2) annotating 100 images with Pleural_Thickening; 3) randomly selecting and annotating 50 images, and 4) randomly selecting and annotating 100 images. Efforts were made to maintain consistency in the total number of labels covered across these strategies. For each, the model was fine-tuned using a combined approach of prediction and attention loss.

*Result & Analysis.* As shown in Table 4, the model's performance varied depending on the annotation strategy. Strategies involving a higher number of annotated images consistently yielded in higher *precision* and *AUC* compared to those using only 50 images. Notably, focusing on Pleural_Thickening, which is underrepresented, led to significant improvements in *recall*. These findings highlight that both annotation *depth* and *breadth* play a crucial role in enhancing the model's generalization capabilities and reducing bias, particularly for tail-end labels.

**Table 3: Performance comparison for *Experiment I: With and Without Attention.***

| Metric | Preliminary Training | Prediction Loss Only | Prediction + Attention Loss |
|---|---|---|---|
| AUC | 0.613 | 0.613 | **0.675** |
| Precision | 0.129 | 0.140 | **0.160** |
| Recall | 0.306 | 0.320 | **0.375** |
| F1 Score | 0.182 | 0.195 | **0.224** |

**Table 4: Performance comparison for *Experiment II: Breadth and Depth.***

| Metric | Preliminary Training | 50 Images (Focused) | 100 Images (Focused) | 50 Images (Random) | 100 Images (Random) |
|---|---|---|---|---|---|
| AUC | 0.613 | 0.620 | **0.675** | 0.610 | 0.631 |
| Precision | 0.129 | 0.138 | **0.160** | 0.131 | 0.141 |
| Recall | 0.306 | 0.312 | **0.375** | 0.295 | 0.323 |
| F1 Score | 0.182 | 0.191 | **0.224** | 0.181 | 0.196 |

## 5.2 User Study

In this section, we compared the experiences of physicians using *MEDebiaser* with their typical interactions with AI models. To validate the effectiveness of our work in real-world scenarios, we chose to conduct our *user study* in the field of otolaryngology, as endoscopic images of the ear typically display multiple distinct pathological features, and there is a certain degree of correlation between different symptoms.

*5.2.1 Participants.* We collaborated with otolaryngology experts from [Blinded for Review], and computer science experts from a local university. We recruited 12 participants, as detailed in Table 5 (physicians: mean age = 40, SD = 6.56; engineers: mean age = 32.7, SD = 4.19). The group included 6 otolaryngology experts and 6 ML experts, with 3 holding a Master's degree, 5 holding an M.D., and 4 holding a Ph.D.. Notably, seven of the participants had prior experience collaborating with counterparts in medical machine learning projects. Three physicians have no background or limited experience with AI. Participants were randomly paired into 6 teams consisting of one physician and one engineer.

*5.2.2 Dataset and Model.* In the *user study*, we utilized an ear endoscopy dataset provided by participants of the study, sourced from [Blinded for Review]. This dataset was collected from 1, 272 patients between 2018 and 2022, comprising a total of 3, 252 images, each with a corresponding diagnostic report. We extracted information from the reports and labeled the associated images. Initially, 12 labels were present in the dataset, but after excluding labels with insufficient representation, we refined it to include 7 labels across 3, 084 images, referred to as **EarEndo**. This dataset exhibits significant issues of *Imbalanced Distribution* and *Label Co-occurrence*, as shown in Figure 11, Appendix E and Table 13, Appendix E. In accordance with the hospital's confidentiality policies, patient names, IDs, addresses, and other personal information were anonymized. Due to the limited dataset size and time constraints of the user study, we instructed the physicians to use *DenseNet* with the default parameter settings (batch size = 4, epoch = 30) throughout the process. The study was conducted on an NVIDIA 4070 Ti GPU.

*5.2.3 Procedure and Task.* Figure 4 illustrates the procedure of our *user study*. Before the study, participants signed privacy and confidentiality agreements and completed a pre-task questionnaire to collect demographic information and task-related background. One author then conducted a 10-minute tutorial on *MEDebiaser* usage and assigned the tasks for the study (Table 6). Then, each team used *MEDebiaser* to mitigate bias related to a specific label. During this period, only physicians operated, while physicians and engineers were restricted from communication except during the *Evaluating Model Performance* stage. The physicians' interactions with *MEDebiaser* were screen recorded, and conversations between physicians and engineers were recorded. Afterwards, participants completed a post-task questionnaire individually. Finally, each team participated in an approximately 20-minute interview. The entire *user study* lasted about one hour on average per participant, who received a $10 token of appreciation upon completion.

We carefully chose tasks based on physicians' typical work pattern (Figure 1) to address key aspects of **RQ4** and **RQ5**, in order to measure both technical analysis and collaborative interaction. This approach allowed us to gain a holistic understanding of *MEDebiaser*'s impact on improving both model performance and the user experience.

*5.2.4 Result Processing and Measurement.* In the screen recordings of interactions with *MEDebiaser*, one author was tasked with capturing user activities, including the number and duration of each fine-tuning session. For the audio recordings of conversations between physicians and engineers, as well as participant interviews, we transcribed these into text and conducted a thematic analysis [61] by coding the transcripts. The time spent on model training and fine-tuning was not included.

**Table 5: The details of *User Study* participants.**

| Physicians - Otolaryngology | | | | | Engineers | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ID** | **Gender** | **Age** | **Experience**[1] | **Degree** | **ID** | **Gender** | **Age** | **Experience**[2] | **Degree** |
| **UD1** | Male | 45 | Yes | M.D. | **UP1** | Female | 29 | Yes | Master |
| **UD2** | Male | 38 | No | M.D. | **UP2** | Male | 32 | Yes | Ph.D. |
| **UD3** | Female | 50 | Yes | Ph.D. | **UP3** | Female | 27 | Yes | Master |
| **UD4** | Male | 42 | Yes | M.D. | **UP4** | Female | 40 | No | Ph.D. |
| **UD5** | Female | 35 | No | M.D. | **UP5** | Male | 35 | No | Master |
| **UD6** | Male | 30 | No | M.D. | **UP6** | Male | 33 | Yes | Ph.D. |

[1] Experience stands for Experience with engineers & AI.
[2] Experience stands for Experience with physicians.

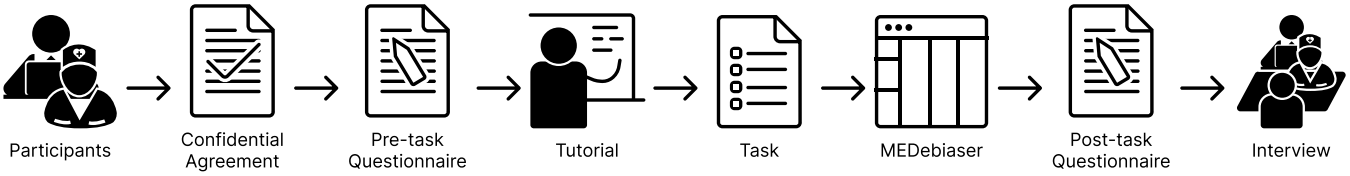


**Figure 4: The procedure of the *User Study*.**

**Table 6: Tasks for Using *MEDebiaser*.**

| ID | View | Parties | Tasks |
|---|---|---|---|
| **T1** | Label View | Physicians | Describe the distribution and co-occurrence of each label in the dataset. (**RQ4**) |
| **T2** | Attention View | Physicians | Identify biased images, particularly those with infrequently occurring labels and label co-occurrence. (**RQ4**) |
| **T3** | Modification View | Physicians | Adjust the attention for images with biased labels. (**RQ4**, **RQ5**) |
| **T4** | Performance View | Physicians & Engineers | Observe and analyze model results after preliminary training and each round of fine-tuning. (**RQ4**, **RQ5**) |

For the post-task questionnaire, a 7-point Likert scale was employed (1: Not at all/Strongly disagree, 7: Very much/Strongly agree). Further details are available in Table 7. The questions were categorized into four aspects: *bias detection*, *system usability scale* [7], *feedback & communication*, and *workload* [34]. Among these, the *system usability scale* and *bias detection* primarily addressed **RQ4**, while *feedback & communication* and *workload* primarily addressed **RQ5**. Even though engineers do not directly interact with the system, their feedback on bias detection, usability, and workload can provide valuable insights into the system's overall performance in a collaborative setting. Thus, as observers, they also filled out the same questionnaire.

For **T1**, the authors rated each participant's description of the dataset on a 5-point Likert scale based on its alignment with the actual data. For **T2** and **T3**, we analyzed the recorded interactions, counting the number of tasks completed, as well as the time and speed taken. For **T4**, we combined the scores from relevant questions in the post-task questionnaire with the analysis of the audio recordings to assess participants' completion of the task and filter valid questionnaires.

## 5.3 User Study Results

We compiled information on users' interactions with *MEDebiaser* through the study. Initially, all physicians successfully uploaded their datasets and selected a model in the *Panel View*. They then proceeded to the *Label View*, where five physicians sorted the table in ascending order by distribution percentage. These physicians selected categories with lower distribution, such as EOM, OME, and EOF, and closely examined the label co-occurrence in the chord diagram. In the *Attention View*, they reviewed 38 images with a total of 49 labels under EOF. Among these images, 18 were correctly predicted, covering 25 labels, while 20 images were incorrectly predicted. Four physicians zoomed in on each image to correct biases

**Table 7: Details of the post-task questionnaire.**

| Aspect | Topic | Question |
|---|---|---|
| Bias Detection (**RQ4**) | Imbalance Mitigation | Does *MEDebiaser* show improvement in mitigating imbalanced label distributions compared to using the model directly? |
| | Co-occurrence Distinction | Does *MEDebiaser* show improvement in distinguishing co-occurring labels compared to using the model directly? |
| System Usability Scale (**RQ4**) | Ease of Learn | How easy was it to learn to use *MEDebiaser*? |
| | Ease of Use | How user-friendly do you find *MEDebiaser*'s interface and interactions? |
| | Likelihood of Future Use | How likely are you to recommend using *MEDebiaser* in the future? |
| Feedback & Communication (**RQ5**) | AI to Physicians | How clear and intuitive is the feedback provided by the AI model to physicians? |
| | Physicians to AI | How effectively can physicians provide human knowledge to improve the AI model? |
| | Between Physicians and Engineers | How well does *MEDebiaser* facilitate collaboration between physicians and engineers? |
| Workload (**RQ5**) | Time Aspect | Does using *MEDebiaser* reduce the time it takes to complete tasks? |
| | Physical Aspect | Does using *MEDebiaser* reduce the physical demands? |

in unreasonable predictions, with **UD3** and **UD4** opting to directly select correctly predicted attention and adjust only for incorrect predictions. In the *Modification View*, all six physicians adeptly used polygons to annotate the images, and **UD1** and **UD5** also utilized the dropdown menu to annotate the remaining labels. For recommendation, **UD1**, **UD2**, and **UD6** chose the dependency mode for sorting, **UD3** and **UD5** selected concentration, and **UD4** opted for accuracy. In the *Performance View*, the physicians reviewed EOF, carefully comparing each image with the results from the previous round. During this stage, the physicians discussed the model's metrics and collaborated with engineers to get further assurance.

*5.3.1 RQ4: Does the system reduce bias in MLMIC and improve model performance, and how do physicians interact with the system to mitigate these biases?* **Using *MEDebiaser*, users gain a better understanding of their datasets, can more easily identify biased images, make necessary adjustments, and ultimately enhance the model's performance.** During the interview, physicians effectively described the overall label distribution in **EarEndo** and the co-occurrence of specific labels (**T1**), with all participants receiving good scores in the rating, offering insights from a medical perspective. Through the *Attention View*, which presents local explanations, physicians reported that they could swiftly identify images with incorrect predictions (**T2**). **UD2** noted, "*The explanations for correctly predicted images usually make sense, but sometimes they aren't detailed [enough] and might need some tweaks. For images that were predicted incorrectly, the explanations are obviously off, so I focus on fixing those [first].*"

Regarding imbalance mitigation (physicians: mean = 5.67; engineers: mean = 4.83) and co-occurrence distinction (physicians: mean = 5.17; engineers: mean = 4.5), it is clear from Figure 5 that physicians rated these aspects higher than engineers. Several physicians observed that, in the *Performance View*, after multiple rounds of fine-tuning, the attention seems to align more accurately with the correct regions. However, two engineers mentioned that the relatively small amount of annotated data in the *user study* has led to only minor changes in metrics. **UP1** commented, "*Right now, with the limited number of annotations, the metrics only show small improvements. But I think as we add more [annotations], we'll start to see a bigger reduction [in bias].*"

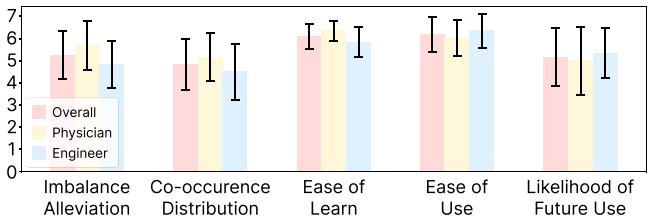


**Figure 5: Likert Results on *Bias Detection* and *System Usability Scale*.**

***MEDebiaser* is easy to use, and its interactive human-AI feedback mechanism effectively aids users in correcting model biases with ease.** Both physicians and engineers acknowledged the ease of use of *MEDebiaser* (physicians: mean = 6; engineers: mean = 6.33) and its low learning curve (physicians: mean = 6.33; engineers: mean = 5.83), as shown in Figure 5. Notably, **UD2**, **UD5**, and **UD6**, despite having no prior ML experience, also found *MEDebiaser* accessible and agreed that it requires minimal ML knowledge to operate. **UD3** mentioned, "*MEDebiaser is quite user-friendly and easy to use. It's similar to Labelme [71], which I have used before for*

*image annotation.*" As shown in Table 8, all six teams completed the modification work for EOF within the designated time (**T3**). Among them, **UD4** completed the task with only three rounds of fine-tuning in the shortest time, while **UD5** took the longest, finishing in five rounds. This difference is attributed to **UD4**'s preference for creating rough outlines with fewer points, compared to **UD5**'s more detailed approach with precise points, as shown in Figure 6. This iterative feedback mechanism allowed physicians to observe changes and trends in the model, facilitating dynamic adjustments and enhancing efficiency. As **UD1** noted, "*After two rounds of adjustments, I could see the results getting better (Figure 7). I was thinking of moving on to other [labels], but I ended up sticking with the remaining [EOF] images as needed.*" Overall, both physicians and engineers expressed a willingness to use *MEDebiaser* in the future (physicians: mean = 5; engineers: mean = 5.33).

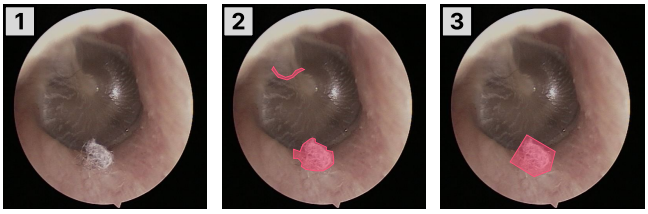


**Figure 6: 1 Original Image. 2 Polygon annotated by UD5. 3 Polygon annotated by UD4.**

*5.3.2 RQ5: What's the impact of the system on workload, physician diagnosis, collaboration, and medical knowledge integration?* ***MEDebiaser*'s interface effectively facilitates iterative feedback between physicians and AI models, promoting the integration of expertise into the training process.** *MEDebiaser* employs Grad-CAM to provide local explanations. Despite being a conventional method, both physicians and engineers found this visualization straightforward and easy to interpret (physicians: mean = 5.67; engineers: mean = 5.83), as shown in Figure 8. **UD6** pointed out, "*We frequently request visualizations, but sometimes they end up being overly complicated and hard to use. Most of the time, simple heatmaps do the job perfectly well.*" Regarding the integration of medical expertise, physicians expressed high satisfaction (mean = 6.33). **UD1** mentioned, "*Previously, it was really challenging for me to communicate the important [medical] features I had in mind to engineers and have them implemented. However, with this new approach, my ideas are now directly and accurately represented in the model, and the heatmaps and line charts it produces are clear and easy to understand.*" (**T2**, **T4**). Engineers found this method of knowledge integration more accessible to physicians and less complex compared to existing approaches (mean = 5.67). **UP3** commented, "*MEDebiaser's interface is designed with clarity, making it easy to interpret the model and guide users effectively. We often use iterative training methods in our work, but usually lack a user interface to facilitate these [processes].*"

**With *MEDebiaser*, the physician-engineer collaboration has shifted from the traditional, lengthy "Bias⇆Revision" cycle to a more dynamic mode where physicians take the lead, with engineers offering support.** Both engineers and physicians generally agree that *MEDebiaser* facilitates their collaboration (physicians: mean = 5.33; engineers: mean = 5.5). **UP5** mentioned, "*This essentially lets doctors make initial adjustments before they bring up any issues to us. If it delivers the results we're aiming for, it definitely makes our job easier.*" **UP2** added, "*This approach helps break [down] some of the communication barriers between us and them. Sometimes we don't completely grasp the specific outcomes they're looking for. Now, they can make initial corrections on their own, without having to depend on us for everything.*" In the traditional mode, physicians often served as AI users, while engineers acted as the AI providers, creating a demand-driven "client-provider" dynamic where engineers held most of the responsibility for managing and maintaining the models. *MEDebiaser* shifts this paradigm, fostering joint management and maintenance of AI by physicians and engineers as equal collaborators. With this tool, engineers are no longer tasked solely with ongoing modifications and adjustments—many of these responsibilities are now shared with physicians in a way that aligns with their expertise. Physicians, in turn, are empowered

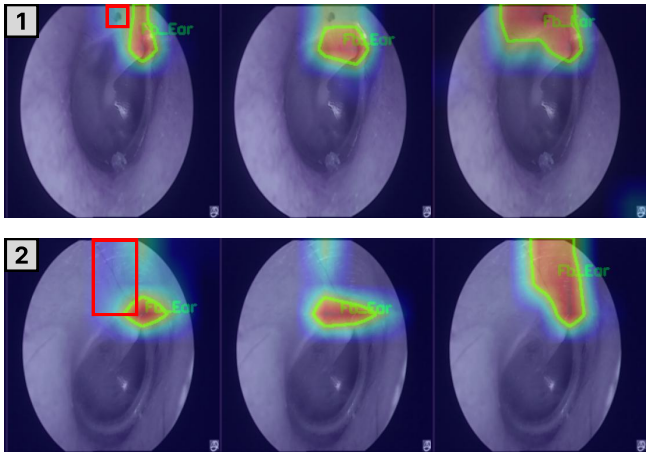


**Figure 7: The results of two rounds of fine-tuning on EOF by UD1. In 1, the red box contains a black foreign object. After each fine-tuning, the attention gets closer to the foreign object. After two rounds of fine-tuning, the attention covers the foreign object. In 2, the red box contains a strand of hair. After each fine-tuning, the attention gets closer to the hair. After two rounds of fine-tuning, the attention covers the entire hair.**

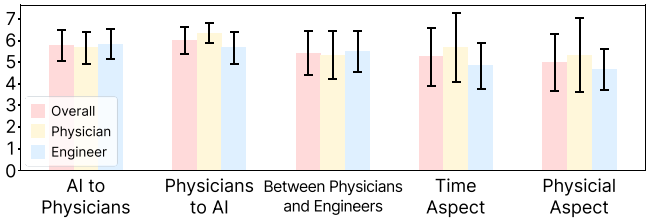


**Figure 8: Likert Results on *Feedback & Communication* and *Workload.***

**Table 8: The details of using *MEDebiaser* on EOF.**

| | Round 1 | | Round 2 | | Round 3 | | Round 4 | | Round 5 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Num.** | **Time** | **Num.** | **Time** | **Num.** | **Time** | **Num.** | **Time** | **Num.** | **Time** | **Num.** | **Time** |
| **UD1** | 8 | 177s | 10 | 184s | 13 | 300s | 9 | 165s | 9 | 200s | 49 | 1,026s |
| **UD2** | 7 | 149s | 5 | 108s | 6 | 124s | 7 | 137s | — | — | 25 | 518s |
| **UD3** | 6 | 138s | 4 | 85s | 5 | 129s | 3 | 67s | — | — | 18 | 419s |
| **UD4** | 5 | 102s | 7 | 155s | 6 | 131s | — | — | — | — | 18 | 388s |
| **UD5** | 12 | 278s | 9 | 213s | 13 | 305s | 8 | 172s | 7 | 159s | 49 | 1,127s |
| **UD6** | 4 | 84s | 6 | 137s | 8 | 188s | 3 | 65s | — | — | 23 | 474s |

**Num.** stands for the number of labels.

to apply their domain knowledge more effectively, without being constrained by the knowledge gap between the two fields. This collaborative mode complements, rather than replaces, the traditional "Bias⇆Revision" cycle. Physicians can still seek assistance from engineers when needed, but this new framework allows both parties to utilize their strengths more efficiently and productively.

***MEDebiaser* has effectively reduced both the time and physical workload for physicians and engineers.** Most physicians found its interactive mechanism significantly decreased the number of required annotations compared to annotating the entire dataset before training, reducing time (mean = 5.67) and physical workload (mean = 5.33). However, **UD2** commented, "*Even though it cuts down on the number of annotations compared to labeling the entire dataset, I still don't want to spend too much [time] on this annotation [work].*" Regarding the physical workload, physicians had varying opinions. **UD4**, who used fewer points for each annotation, felt the physical workload was manageable, while **UD5**, who meticulously annotated each image, expressed concern: "*When there's a lot of data, this work can get exhausting, both physically and mentally.*" engineers also reported reduced time workload (mean = 4.83) and physical workload (mean = 4.67). **UP1** noted, "*Our main role is to guide the training settings and confirm the results, which is a lot easier than having to fix the model whenever something goes wrong, especially when we don't fully understand the data.* "

## 6 Discussion

### 6.1 Design Implications

*6.1.1 Division of Responsibilities Between Physicians and Engineers.* In complex medical human-AI collaboration, clearly defining the roles of physicians and engineers is crucial for effective outcomes. In *MEDebiaser*, physicians go beyond annotating data by interpreting model outputs and understanding how their feedback influences training, as noted by **UD2**: "*It's important for me to not only provide feedback but also understand how it impacts the model's behavior over time.*" Meanwhile, engineers handle the technical implementation, ensuring model performance and explaining its behavior to physicians. They maintain control over the model's integrity and adjust algorithms as needed. Regardless of the specific collaboration mode, the key principle is aligning roles with each party's expertise, ensuring continuous feedback, and reducing dependency. Such collaboration leads to more effective human-AI solutions. Future systems should integrate this to maximize both human expertise and AI capabilities for improved healthcare outcomes.

*6.1.2 Human Workload, AI Performance, and Their Tradeoff.* In *MEDebiaser* and similar AI systems, the human workload is a critical factor in determining the system's scalability. Unlike traditional annotation methods that require labeling the entire dataset, *MEDebiaser* reduces this burden by focusing specifically on biased images and offering customized recommendation strategy to further ease the workload. In the *user study*, **UD5** raised concerns about the increasing manual annotation efforts required for improving model accuracy and mitigating bias. These concerns are valid, as manual annotation can be taxing for physicians, especially with large datasets. However, our work targets the underrepresented tail-end labels that contribute to model bias, aiming to enhance model fairness and performance without overwhelming users with excessive manual labor.

Another challenge in the medical domain, unlike natural image classification, is the scarcity of data for rare classes, where features are often sparse and difficult to capture, even with state-of-the-art models. While no tool can fully compensate for the lack of data, the most effective solution is to increase the number of cases. *MEDebiaser* works within the real-world data constraints, leveraging physician involvement to help the model focus on rare classes. The data distribution used in our experiments aligns with that in most research on imbalanced datasets, and both the *mechanism study* and *user study* validate the effectiveness of our approach. However, it is important to acknowledge that the limitations of sparse medical data will still impact model performance, and our approach is an effective attempt to partially mitigate this issue.

In the *MEDebiaser* framework, while annotating additional data enhances model performance, it may also increase the workload for physicians. To mitigate this, our system employs a customized ranking strategy that prioritizes the most critical and underrepresented labels, thereby reducing the amount of annotation required. This approach helps minimize manual effort while improving both model accuracy and fairness. However, there is potential for further optimization by incorporating more advanced automation techniques, such as diffusion learning [62]. Diffusion learning can facilitate the transfer of patterns from annotated to unannotated images, enabling the model to automatically infer labels and reduce the need for manual annotations [62]. Physicians could then review

and refine these auto-generated annotations, further alleviating their workload. Similar strategies have been successfully applied in medical text classification tools [51, 85], where expert knowledge and machine learning work collaboratively to improve efficiency. However, the use of such technologies requires significant design and experimentation to ensure their effectiveness and safety. Therefore, we have not explored or attempted this approach further in our work.

*6.1.3 Human Autonomy, AI Controllability, and Their Tradeoff.* In the *user study*, we observed that some participants, such as **UD4**, exhibited behaviors indicative of a group that prefers to delegate much of the task to AI, especially for tasks like image segmentation and detection, which they believe are already highly mature. **UD4** completed tasks quickly, with fewer annotations and polygon vertices. This behavior, however, represents one end of a spectrum of interaction styles we observed. In stark contrast, other physicians adopted a more meticulous and cautious approach, creating highly detailed annotations with numerous vertices, reflecting a desire to maintain greater control over the process. This potential variability in physician annotation behavior, stemming from differing levels of trust in AI and personal diligence, is a critical factor to consider. The sentiment from **UD4**, who mentioned during the interview, "*AI is already powerful enough, just let it handle the task,*" exemplifies the "delegator" attitude and reflects a common inclination in medical settings to over-rely on AI for tasks considered well-suited to automation. However, while the potential for automation in healthcare is evident, its feasibility largely depends on the accuracy of the model, as a fully automated system cannot be guaranteed to achieve 100% accuracy.

While AI systems have demonstrated remarkable performance in tasks like medical image segmentation and classification, it is crucial to recognize that these models' effectiveness is often dependent on the data distribution used during training. Consequently, these models may struggle when confronted with rare or atypical cases. This highlights the continuing need for human involvement. Physicians should not be expected to compete with AI in tasks where it excels, such as image segmentation or classification. However, their expertise is indispensable in areas where AI may falter—such as interpreting ambiguous findings, incorporating patient history, and identifying subtle cues beyond the model's scope. In these instances, human oversight ensures AI's performance aligns with real-world clinical requirements, particularly in handling rare or minority cases that may be misclassified due to data imbalance. Therefore, even in highly automated settings, human intervention remains necessary to validate and guide AI-driven decisions.

Currently, *MEDebiaser* offers visual feedback on intermediate results, providing some transparency into the model's behavior. However, this doesn't fully ensure understanding or control, as physicians may struggle to interpret the data and determine the right parameter adjustments, potentially disrupting the training process. To address this, future versions could include an automatic recommendation system that suggests optimal parameter settings based on training progress, as well as anomaly detection and real-time alerts to notify physicians of irregularities. These enhancements would enable physicians to monitor and control the model more effectively, ensuring that training remains on track and that optimal adjustments are made.

## 6.2 Generalizability

*6.2.1 Data Update.* In the formative study, **D3** remarked: "*We see a large number of patients every day, generating new images and diagnoses. It would be ideal if this data could be utilized.*" The continuous influx of new data presents a challenge in keeping the model up-to-date and adaptive to evolving trends. *MEDebiaser* addresses this by recommending the most valuable data points for physicians to annotate, particularly when new data becomes available. By identifying uncertain or unfamiliar cases, the system minimizes effort spent on well-understood samples and instead prioritizes those that contribute most to model improvement. This targeted recommendation streamlines the annotation process while ensuring the model remains aligned with the latest medical knowledge, enhancing its generalizability over time. Consequently, *MEDebiaser* is well-suited for deployment in hospital environments where data is continuously updated.

*6.2.2 Broader Applications in Medical and Non-Medical Scenarios.* In real-world scenarios, multi-label data is common. *MEDebiaser* has shown effectiveness in reducing bias in such datasets, making it suitable for broader applications across various complex, multi-label contexts. What sets *MEDebiaser* apart from other state-of-the-art MLMIC models is its interactive design, allowing integration of human expertise during training, unlike models where physicians have to passively accept outputs. This interactivity boosts its utility in medical contexts and beyond. While our evaluation focuses on two datasets, *MEDebiaser* is applicable to other medical multi-label tasks because its core workflow is designed to address fundamental challenges, like imbalanced distributions and label co-occurrence, that are common across many medical imaging domains, such as the analysis of skin lesions in dermatology, tissue classification in digital pathology, and the identification of multiple findings in brain MRI scans. Additionally, while medical annotations depend on domain experts, natural image datasets typically do not require such specialized expertise for annotations. In these cases, crowdsourcing [36] can harness public input, and when combined with AI systems like *MEDebiaser*, generate accurate and reliable labels even without specialized knowledge.

## 6.3 Limitation & Future Work

*6.3.1 Recommendation Strategy.* The current recommendation strategy in *MEDebiaser* uses a ranking method based on label accuracy, heatmap concentration, and co-occurrence matrix dependency, focusing on model-driven metrics over medical ones. **UD3** noted, "*The system's recommendations are helpful, but they don't always match what we prioritize in clinical practice. It would be better if it could suggest cases based on medical similarities rather than just relying on model outputs.*" Physicians prefer recommendations based on medical feature similarity to better analyze specific symptoms. Future work could develop a dynamic recommendation system grounded in medical features and tailored to physicians' annotation patterns, helping them identify model issues more effectively. Additionally, recommendations could be tailored to physicians' actual annotation patterns. For example, if a certain region is frequently annotated, it

may suggest that symptoms in that area are more challenging to recognize or prone to model misclassification. In such cases, the system could prioritize recommending images where the Grad-CAM heatmap does not highlight that region, thereby helping physicians more efficiently identify potential model issues and areas for improvement.

*6.3.2 Practical Deployment Considerations.* Moving *MEDebiaser* from a research prototype to a clinical tool involves several practical considerations. For data privacy, the system is designed for on-premises deployment within a hospital's secure network to comply with strict regulations. Regarding hardware, while the front-end is a lightweight web application, the back-end requires a server with a modern GPU for efficient model fine-tuning. Physician training must be brief and practical, and the tool must be integrated into clinical workflows as a retrospective activity requiring dedicated time. Beyond these, a clear protocol for model maintenance and versioning is crucial; engineers must validate each new physician-tuned model against a hold-out dataset before it is promoted for wider use. Finally, for long-term adoption, future work should address interoperability with existing hospital systems, such as integrating with Picture Archiving and Communication System (PACS) to streamline image selection and review, further embedding the tool into the natural clinical workflow.

## 7 Conclusion

This study presents *MEDebiaser*, an interactive system designed to mitigate bias in MLMIC. *MEDebiaser* enhances human-AI collaboration by facilitating continuous feedback between physicians and AI models, offering interpretability, and enabling physicians to directly adjust the model's attention during iterative fine-tuning through an intuitive interface. At the same time, it ensures effective co-supervision of AI models between physicians and engineers. By streamlining the physician-engineer collaboration process, *MEDebiaser* reduces the workload for both parties. Our *mechanism study* and *user study* demonstrate that *MEDebiaser* significantly reduces bias in MLMIC, improves usability, and boosts the overall efficiency of physician-engineer collaboration.

## References


[1] Saranya A. and Subhashini R. 2023. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal* 7 (2023), 100230. https://doi.org/10.1016/j.dajour.2023.100230

[2] A.S. Albahri, Ali M. Duhaim, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* 96 (2023), 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

[3] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 232 (jan 2021), 31 pages. https://doi.org/10.1145/3432931

[4] Markus Bertl, Toomas Klementi, Gunnar Piho, Peeter Ross, and Dirk Draheim. 2023. How Domain Engineering Can Help to Raise Adoption Rates of Artificial Intelligence in Healthcare. In *Information Integration and Web Intelligence*, Pari Delir Haghighi, Eric Pardede, Gillian Dobbie, Vithya Yogarajan, Ngurah Agus Sanjaya ER, Gabriele Kotsis, and Ismail Khalil (Eds.). Springer Nature Switzerland, Cham, 3–12. https://doi.org/10.1007/978-3-031-48316-5_1

[5] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M. Friedrich, and Felix Nensa. 2023. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. *European Journal of Radiology* 162 (2023), 110786. https://doi.org/10.1016/j.ejrad.2023.110786

[6] Serdar Bozyel, Evrim Şimşek, Duygu Koçyiğit Burunkaya, Arda Güler, Yetkin Korkmaz, Mehmet Şeker, Mehmet Ertürk, and Nurgül Keser. 2024. Artificial Intelligence-Based Clinical Decision Support Systems in Cardiovascular Diseases. *Anatolian Journal of Cardiology* 28, 2 (January 7 2024), 74–86. https://doi.org/10.14744/AnatolJCardiol.2023.3685 PMID: 38168009.

[7] John Brooke. 2013. SUS: a retrospective. *J. Usability Studies* 8, 2 (feb 2013), 29–40.

[8] Francisco Maria Calisto, João Maria Abrantes, Carlos Santiago, Nuno J. Nunes, and Jacinto C. Nascimento. 2025. Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. *International Journal of Human-Computer Studies* 197 (2025), 103444. https://doi.org/10.1016/j.ijhcs.2025.103444

[9] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 13, 20 pages. https://doi.org/10.1145/3544548.3580682

[10] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (2022), 102922. https://doi.org/10.1016/j.ijhcs.2022.102922

[11] Yidong Chai, Hongyan Liu, Jie Xu, Sagar Samtani, Yuanchun Jiang, and Haoxin Liu. 2023. A Multi-Label Classification with an Adversarial-Based Denoising Autoencoder for Medical Image Annotation. *ACM Trans. Manage. Inf. Syst.* 14, 2, Article 19 (jan 2023), 21 pages. https://doi.org/10.1145/3561653

[12] Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. 2020. Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification. *IEEE Journal of Biomedical and Health Informatics* 24, 8 (2020), 2292–2302. https://doi.org/10.1109/JBHI.2020.2967084

[13] Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. 2021. OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples. *IEEE Transactions on Visualization and Computer Graphics* 27, 7 (July 2021), 3335–3349. https://doi.org/10.1109/TVCG.2020.2973258

[14] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine* 5, 1 (2022), 156. https://doi.org/10.1038/s41746-022-00699-2

[15] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. 2019. Learning Semantic-Specific Graph Representation for Multi-Label Image Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 522–531. https://doi.org/10.1109/ICCV.2019.00061

[16] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5172–5181. https://doi.org/10.1109/CVPR.2019.00532

[17] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2023. Learning Graph Convolutional Networks for Multi-Label Recognition and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 6969–6983. https://doi.org/10.1109/TPAMI.2021.3063496

[18] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. https://doi.org/10.48550/arXiv.1810.08810 arXiv:1810.08810 [cs.LG]

[19] Haluk Demirkan and Dursun Delen. 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decis. Support Syst.* 55, 1 (apr 2013), 412–421. https://doi.org/10.1016/j.dss.2012.05.048

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[21] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S. Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3099–3102. https://doi.org/10.1145/2556288.2557011

[22] Joseph Donia and James A. Shaw. 2021. Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Society* 8, 2 (2021), 20539517211065248. https://doi.org/10.1177/20539517211065248

[23] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. https://doi.org/10.1145/3185517

[24] Kevin Figueroa, Bofan Song, Sumsum Sunny, Shaobai Li, Keerthi Gurushanth, Pramila Mendonca, Nirza Mukhia, Sanjana Patrick, Shubha Gurudath, Subhashini Raghavan, Imchen Tsusennaro, Shirley T. Leivon, Trupti Kolur, Vivek

Shetty, Vidya Bushan, Rohan M. Ramesh, Vijay Pillai, Petra Wilder-Smith, Alben Sigamani, Amritha Suresh, Moni A. Kuriakose, Praveen Birur, and Rongguang Liang. 2022. Interpretable deep learning approach for oral cancer classification using guided attention inference network. *Journal of Biomedical Optics* 27, 1 (2022), 015001. https://doi.org/10.1117/1.JBO.27.1.015001

[25] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10697–10706. https://doi.org/10.1109/CVPR.2019.01096

[26] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2024. Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning. *ACM Comput. Surv.* 56, 7, Article 188 (apr 2024), 39 pages. https://doi.org/10.1145/3644073

[27] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 489 (nov 2022), 28 pages. https://doi.org/10.1145/3555590

[28] Shizhan Gong, Cheng Chen, Yuqi Gong, Nga Yan Chan, Wenao Ma, Calvin Hoi-Kwan Mak, Jill Abrigo, and Qi Dou. 2023. Diffusion model based semi-supervised learning on brain hemorrhage images for efficient midline shift quantification. In *International Conference on Information Processing in Medical Imaging*. Springer, 69–81. https://doi.org/10.1007/978-3-031-34048-2_6

[29] Liang Gou, Lincan Zou, Nanxiang Li, Michael Hofmann, Arvind Kumar Shekar, Axel Wendt, and Liu Ren. 2021. VATLD: A Visual Analytics System to Assess, Understand and Improve Traffic Light Detection. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 261–271. https://doi.org/10.1109/TVCG.2020.3030350

[30] Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, Cara Calvelli, Pengcheng Shi, and Anne Haake. 2016. An Expert-in-the-loop Paradigm for Learning Medical Image Grouping. In *Advances in Knowledge Discovery and Data Mining*, James Bailey, Latifur Khan, Takashi Washio, Gill Dobbie, Joshua Zhexue Huang, and Ruili Wang (Eds.). Springer International Publishing, Cham, 477–488. https://doi.org/10.1007/978-3-319-31753-3_38

[31] Shivam Gupta, Sachin Modgil, Samadrita Bhattacharyya, and Indranil Bose. 2021. Artificial intelligence for decision support systems in the field of operations research: review and future scope of research. *Annals of Operations Research* 308 (2021), 215 – 274. https://doi.org/10.1007/s10479-020-03856-6

[32] Meng Han, Hongxin Wu, Zhiqiang Chen, Muhan Li, and Xilong Zhang. 2022. A survey of multi-label classification based on supervised and semi-supervised learning. *International Journal of Machine Learning and Cybernetics* 14 (2022), 697–724. https://doi.org/10.1007/s13042-022-01658-9

[33] Allan Hanbury, Henning Mller, and Georg Langs. 2017. *Cloud-Based Benchmarking of Medical Image Analysis* (1st ed.). Springer Publishing Company, Incorporated. https://doi.org/10.1007/978-3-319-49644-3

[34] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[36] Eric Heim, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, Alexander W. Sauter, Fides Regina Schwartz, Andreas Termer, Felix Wagner, Hannes Götz Kenngott, and Lena Maier-Hein. 2018. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging* 5, 3 (2018), 034002. https://doi.org/10.1117/1.JMI.5.3.034002

[37] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 793–811. https://doi.org/10.1007/978-3-030-01219-9_47

[38] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300809

[39] Gregory Holste, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, Dongkyun Kim, Trong-Hieu Nguyen-Mau, Minh-Triet Tran, Jaehyup Jeong, Wongi Park, Jongbin Ryu, Feng Hong, Arsh Verma, Yosuke Yamagishi, Changhyun Kim, Hyeryeong Seo, Myungjoo Kang, Leo Anthony Celi, Zhiyong Lu, Ronald M. Summers, George Shih, Zhangyang Wang, and Yifan Peng. 2024. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Medical Image Analysis* 97 (2024), 103224. https://doi.org/10.1016/j.media.2024.103224

[40] Feng Hong, Tianjie Dai, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2023. Bag of Tricks for Long-Tailed Multi-Label Classification on Chest X-Rays. https://doi.org/10.48550/arXiv.2308.08853 arXiv:2308.08853 [cs.CV]

[41] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[42] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[43] Jaehyup Jeong, Bosoung Jeoun, Yeonju Park, and Bohyung Han. 2023. An Optimized Ensemble Framework for Multi-Label Classification on Long-Tailed Chest X-ray Data. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2731–2738. https://doi.org/10.1109/ICCVW60793.2023.00289

[44] Liu Jiang, Shixia Liu, and Changjian Chen. 2019. Recent research advances on interactive machine learning. *J. Vis.* 22, 2 (apr 2019), 401–417. https://doi.org/10.1007/s12650-018-0531-1

[45] Mohamed Khalifa and Mona Albadawy. 2024. AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update* 5 (2024), 100146. https://doi.org/10.1016/j.cmpbup.2024.100146

[46] Shahzeb Khan and Jawwad Ahmed Shamsi. 2021. Health Quest: A generalized clinical decision support system with multi-label classification. *Journal of King Saud University - Computer and Information Sciences* 33, 1 (2021), 45–53. https://doi.org/10.1016/j.jksuci.2018.11.003

[47] Changhyun Kim, Giyeol Kim, Sooyoung Yang, Hyunsu Kim, Sangyool Lee, and Hansu Cho. 2023. Chest X-Ray Feature Pyramid Sum Model with Diseased Area Data Augmentation Method. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2749–2758. https://doi.org/10.1109/ICCVW60793.2023.00291

[48] Dongkyun Kim. 2023. CheXFusion: Effective Fusion of Multi-View Features using Transformers for Long-Tailed Chest X-Ray Classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2694–2702. https://doi.org/10.1109/ICCVW60793.2023.00285

[49] Dajung Kim, Niko Vegt, Valentijn Visch, and Marina Bos-De Vos. 2024. How Much Decision Power Should (A)I Have?: Investigating Patients' Preferences Towards AI Autonomy in Healthcare Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 439, 17 pages. https://doi.org/10.1145/3613904.3642883

[50] Qi Lai, Jianhang Zhou, Yanfen Gan, Chi-Man Vong, and C.L. Philip Chen. 2024. Single-Stage Broad Multi-Instance Multi-Label Learning (BMIML) With Diverse Inter-Correlations and Its Application to Medical Image Classification. *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, 1 (2024), 828–839. https://doi.org/10.1109/TETCI.2023.3287978

[51] Xiang Li, Menglin Cui, Jingpeng Li, Ruibin Bai, Zheng Lu, and Uwe Aickelin. 2021. A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing* 443 (2021), 345–355. https://doi.org/10.1016/j.neucom.2021.02.069

[52] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. 2022. The Emerging Trends of Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 7955–7974. https://doi.org/10.1109/TPAMI.2021.3119334

[53] Octavio Loyola-González, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Milton García-Borroto. 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomput.* 175, PB (jan 2016), 935–947. https://doi.org/10.1016/j.neucom.2015.04.120

[54] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[55] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 195 (nov 2019), 30 pages. https://doi.org/10.1145/3359297

[56] T. Martín-Noguerol, F. Paulano-Godino, R. López-Ortega, J.M. Górriz, R.F. Riascos, and A. Luna. 2021. Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. *Clinical Radiology* 76, 5 (2021), 317–324. https://doi.org/10.1016/j.crad.2020.11.113

[57] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (2018), 1236–1246. https://doi.org/10.1093/bib/bbx044 arXiv:PMC6455466 PMID: 28481991.

[58] Eka Miranda, Mediana Aryuni, and E. Irwansyah. 2016. A survey of medical image classification techniques. In *2016 International Conference on Information Management and Technology (ICIMTech)*. 56–61. https://doi.org/10.1109/

ICIMTech.2016.7930302
[59] Elham Nasarian, Roohallah Alizadehsani, U.Rajendra Acharya, and Kwok-Leung Tsui. 2024. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information Fusion* 108 (2024), 102412. https://doi.org/10.1016/j.inffus.2024.102412
[60] Trong-Hieu Nguyen-Mau, Tuan-Luc Huynh, Thanh-Danh Le, Hai-Dang Nguyen, and Minh-Triet Tran. 2023. Advanced Augmentation and Ensemble Approaches for Classifying Long-Tailed Multi-Label Chest X-Rays. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2721–2730. https://doi.org/10.1109/ICCVW60793.2023.00288
[61] Julianne S. Oktay. 2012. *Grounded Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753697.001.0001
[62] Pedro Osorio, Guillermo Jimenez-Perez, Javier Montalt-Tordera, Jens Hooge, Guillem Duran-Ballester, Shivam Singh, Moritz Radbruch, Ute Bach, Sabrina Schroeder, Krystyna Siudak, Julia Vienenkoetter, Bettina Lawrenz, and Sadegh Mohammadi. 2024. Latent Diffusion Models with Image-Derived Annotations for Enhanced AI-Assisted Cancer Diagnosis in Histopathology. *Diagnostics* 14, 13 (2024). https://doi.org/10.3390/diagnostics14131442
[63] Yang Ouyang, Yuchen Wu, He Wang, Chenyang Zhang, Furui Cheng, Chang Jiang, Lixia Jin, Yuanwu Cao, and Quan Li. 2024. Leveraging Historical Medical Records as a Proxy via Multimodal Modeling and Visualization to Enrich Medical Diagnostic Learning. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 1238–1248. https://doi.org/10.1109/TVCG.2023.3326929
[64] Yang Ouyang, Chenyang Zhang, He Wang, Tianle Ma, Chang Jiang, Yuheng Yan, Zuoqin Yan, Xiaojuan Ma, Chuhan Shi, and Quan Li. 2024. A Two-Phase Visualization System for Continuous Human-AI Collaboration in Sequelae Analysis and Modeling. https://doi.org/10.48550/arXiv.2407.14769 arXiv:2407.14769 [cs.HC]
[65] Meghana Padmanabhan, Pengyu Yuan, Govind Chada, and Hien Van Nguyen. 2019. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. *Journal of Clinical Medicine* 8, 7 (2019). https://doi.org/10.3390/jcm8071050
[66] Wongi Park, Inhyuk Park, Sungeun Kim, and Jong Bin Ryu. 2023. Robust Asymmetric Loss for Multi-Label Long-Tailed Learning. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2023), 2703–2712. https://doi.org/10.48550/arXiv.2308.05542
[67] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A Survey on Biomedical Image Captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, Raffaella Bernardi, Raquel Fernandez, Spandana Gella, Kushal Kafle, Christopher Kanan, Stefan Lee, and Moin Nabi (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 26–36. https://doi.org/10.18653/v1/W19-1803
[68] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Comput. Surv.* 54, 9, Article 180 (Oct. 2021), 40 pages. https://doi.org/10.1145/3472291
[69] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, John DeNero, Mark Finlayson, and Sravana Reddy (Eds.). Association for Computational Linguistics, San Diego, California, 97–101. https://doi.org/10.18653/v1/N16-3020
[70] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 82–91. https://doi.org/10.1109/ICCV48922.2021.00015
[71] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision* 77 (2008), 157–173. https://doi.org/10.1007/s11263-007-0090-8
[72] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74
[73] HyeRyeong Seo, MinHyuk Lee, WooJin Cheong, HyeKyung Yoon, SoHyung Kim, and MyungJoo Kang. 2023. Enhancing Multi-Label Long-Tailed Classification on Chest X-Rays through ML-GCN Augmentation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2739–2748. https://doi.org/10.1109/ICCVW60793.2023.00290
[74] Chuhan Shi, Yicheng Hu, Shenan Wang, Shuai Ma, Chengbo Zheng, Xiaojuan Ma, and Qiong Luo. 2023. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 770, 20 pages. https://doi.org/10.1145/3544548.3581469
[75] Wenqi Shi, Li Tong, Yuanda Zhu, and May D. Wang. 2021. COVID-19 Automatic Diagnosis With Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks. *IEEE Journal of Biomedical and Health Informatics* 25, 7 (2021), 2376–2387. https://doi.org/10.1109/JBHI.2021.3074893
[76] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2018), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063
[77] Aram Siamak, Roozbeh Sadeghian, Iheb Abdellatif, and Stanley Nwoji. 2019. Diagnosing Heart Disease Types from Chest X-Rays Using a Deep Learning Approach. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. 910–913. https://doi.org/10.1109/CSCI49370.2019.00173
[78] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. arXiv:1412.6806 [cs.LG] https://arxiv.org/abs/1412.6806
[79] George Sun and Yi-Hui Zhou. 2023. AI in healthcare: navigating opportunities and challenges in digital communication. *Frontiers in Digital Health* 5 (2023), 1291132. https://doi.org/10.3389/fdgth.2023.1291132
[80] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine* 3, 1 (February 2020), 17. https://doi.org/10.1038/s41746-020-0221-y
[81] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2015. A Deeper Look at Dataset Bias. *arXiv e-prints*, Article arXiv:1505.01257 (May 2015), arXiv:1505.01257 pages. https://doi.org/10.48550/arXiv.1505.01257 arXiv:1505.01257 [cs.CV]
[82] A. Torralba and A. A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*. IEEE Computer Society, USA, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347
[83] Arsh Verma. 2023. How Can We Tame the Long-Tail of Chest X-ray Datasets? https://doi.org/10.48550/arXiv.2309.04293 arXiv:2309.04293 [eess.IV]
[84] Guoli Wang, Pingping Wang, and Benzheng Wei. 2024. Multi-label local awareness and global co-occurrence priori learning improve chest X-ray classification. *Multim. Syst.* 30 (2024), 132. https://doi.org/10.1007/s00530-024-01321-z
[85] He Wang, Yang Ouyang, Yuchen Wu, Chang Jiang, Lixia Jin, Yuanwu Cao, and Quan Li. 2024. KMTLabeler: An Interactive Knowledge-Assisted Labeling Tool for Medical Text Classification. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–18. https://doi.org/10.1109/TVCG.2024.3406387
[86] Kai Wang, Shuqi He, Wenlu Wang, Jinbei Yu, Yu Liu, and Lingyun Yu. 2024. CHORDination: Evaluating Visual Design Choices in Chord Diagrams for Network Data. In *Proceedings of the 17th International Symposium on Visual Information Communication and Interaction (VINCI '24)*. Association for Computing Machinery, New York, NY, USA, Article 15, 8 pages. https://doi.org/10.1145/3678698.3678707
[87] Shuo Wang, Leandro L. Minku, and Xin Yao. 2015. Resampling-Based Ensemble Methods for Online Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering* 27, 5 (2015), 1356–1368. https://doi.org/10.1109/TKDE.2014.2345380
[88] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. 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)*. 3462–3471. https://doi.org/10.1109/CVPR.2017.369
[89] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 162–178. https://doi.org/10.1007/978-3-030-58548-8_10
[90] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376807
[91] Yosuke Yamagishi and Shohei Hanaoka. 2023. Effect of Stage Training for Long-Tailed Multi-Label Image Classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2713–2720. https://doi.org/10.1109/ICCVW60793.2023.00287
[92] Weikai Yang, Yukai Guo, Jing Wu, Zheng Wang, Lan-Zhe Guo, Yu-Feng Li, and Shixia Liu. 2024. Interactive Reweighting for Mitigating Label Quality Issues. *IEEE Transactions on Visualization and Computer Graphics* 30, 3 (2024), 1837–1852. https://doi.org/10.1109/TVCG.2023.3345340
[93] Weikai Yang, Mengchen Liu, Zheng Wang, and Shixia Liu. 2024. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media* 10, 3 (2024), 399–424. https://doi.org/10.1007/s41095-023-0393-x

[94] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-Driven Dynamic Graph Convolutional Network for Multi-label Image Recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 649–665. https://doi.org/10.1007/978-3-030-58589-1_39

[95] Seid Muhie Yimam, Chris Biemann, Ljiljana Majnaric, Šefket Šabanović, and Andreas Holzinger. 2015. Interactive and Iterative Annotation for Biomedical Entity Recognition. In *Brain Informatics and Health*, Yike Guo, Karl Friston, Faisal Aldo, Sean Hill, and Hanchuan Peng (Eds.). Springer International Publishing, Cham, 347–357. https://doi.org/10.1007/978-3-319-23344-4_34

[96] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 248, 15 pages. https://doi.org/10.1145/3544548.3580945

[97] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 7, 1 (Mar 2021), 3–36. https://doi.org/10.1007/s41095-020-0191-7

[98] Alwin Yaoxian Zhang, Sean Shao Wei Lam, Nan Liu, Yan Pang, Ling Ling Chan, and Phua Hwee Tang. 2018. Development of a Radiology Decision Support System for the Classification of MRI Brain Scans. In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*. 107–115. https://doi.org/10.1109/BDCAT.2018.00021

[99] Jie Zhang and Zong-ming Zhang. 2023. Ethics and governance of trustworthy medical artificial intelligence. *BMC medical informatics and decision making* 23, 1 (2023), 7. https://doi.org/10.1186/s12911-023-02103-9

[100] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

[101] Xiaoyu Zhang, Xiwei Xuan, Alden Dima, Thurston Sexton, and Kwan-Liu Ma. 2023. LabelVizier: Interactive Validation and Relabeling for Technical Text Annotations. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*. 167–176. https://doi.org/10.1109/PacificVis56936.2023.00026

[102] Yu Zhang, Jing Chen, Xiangxun Ma, Gang Wang, Uzair Aslam Bhatti, and Mengxing Huang. 2024. Interactive medical image annotation using improved Attention U-net with compound geodesic distance. *Expert Systems with Applications* 237 (2024), 121282. https://doi.org/10.1016/j.eswa.2023.121282

[103] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. 2023. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1977–1987. https://doi.org/10.1109/ICCV51070.2023.00189

[104] Yuhan Zhang, Luyang Luo, Qi Dou, and Pheng-Ann Heng. 2023. Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. *Medical Image Analysis* 86 (2023), 102772. https://doi.org/10.1016/j.media.2023.102772

[105] Xuehan Zhao, Jiaqi Liu, Zhiwen Yu, and Bin Guo. 2024. HADT: Human-AI Diagnostic Team via Hierarchical Reinforcement Learning. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. 860–868. https://doi.org/10.1137/1.9781611978032.98

[106] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. 2024. Understanding Nonlinear Collaboration between Human and AI Agents: A Co-design Framework for Creative Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 170, 16 pages. https://doi.org/10.1145/3613904.3642812

[107] Oren Zuckerman, Viva Sarah Press, Ehud Barda, Benny Megidish, and Hadas Erel. 2022. Tangible Collaboration: A Human-Centered Approach for Sharing Control With an Actuated-Interface. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 507, 13 pages. https://doi.org/10.1145/3491102.3517449

## A Content Analysis Supplementary

See Table 9 and Table 10.

## B Loading Dataset and Model Supplementary

See Figure 9 and Table 11.

## C Observing and Modifying Attention Supplementary

See Figure 10.

## D Evaluation Metrics Supplementary

$$precision = \frac{TP}{TP + FP} \tag{3}$$

where True Positive (*TP*) is the number of positive instances correctly predicted as positive; False Positive (*FP*) is the number of negative instances incorrectly predicted as positive.

$$recall = \frac{TP}{TP + FN} \tag{4}$$

where False Negative (*FN*) is the number of positive instances incorrectly predicted as negative.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

$$AUC = \int_0^1 TPR\, d(FPR) \tag{6}$$

where True Positive Rate (*TPR*) is equivalent to Recall; False Positive Rate (*FPR*) is:

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

where True Negative (*TN*) is the number of negative instances correctly predicted as negative.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$specificity = \frac{TN}{FP + TN} \tag{9}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{10}$$

where $N$ is the total number of classes; $AP_i$ is the *average precision* for the $i$-th class:

$$AP_i = \frac{1}{|R_i|} \sum_{r \in R_i} P(r) \cdot \Delta r \tag{11}$$

where $R_i$ is the set of retrieved results for the $i$-th class, ordered by their confidence scores; $P(r)$ is the *precision* at the $r$-th rank; $\Delta r$ is the change in *recall* from the $(r-1)$-th to the $r$-th retrieved result.

See Table 12.

## E User Study Supplementary

See Figure 11 and Table 13.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

**Table 9: Summary of Common Annotation Methods in Medical AI Tools**

| Annotation Method | Description | Example | Advantages | Disadvantages |
|---|---|---|---|---|
| Bounding Boxes | A rectangular box drawn around the object. |  | Simple to implement; widely supported | Inaccurate for irregular anatomical regions |
| Brush / Flood Fill | Filling a region with color to define an area. |  | Intuitive interaction; pixel-level detail | May produce inconsistent results |
| Polygon | A multi-point closed shape drawn around an object. |  | High precision for complex contours | Time-consuming for manual annotation |
| Keypoint Skeleton | Marking specific key points and connecting them with lines. |  | Suitable for landmark/pose annotation | Not applicable to region segmentation |
| Polyline | A series of connected straight lines defining an object. |  | Useful for vessel/trachea tracing | Limited to linear structures |

**Table 10: Results of Usability Evaluation on Annotation Methods**

| Annotation Method | Avg. Time (s) | Accuracy (0–1) | Satisfaction (1–5) |
|---|---|---|---|
| Bounding Boxes | **12.4** | 0.78 | 3.2 |
| Polyline | 15.6 | 0.71 | 3.0 |
| Polygon | 19.7 | **0.92** | **4.5** |
| Brush / Flood Fill | 20.5 | 0.89 | 4.0 |

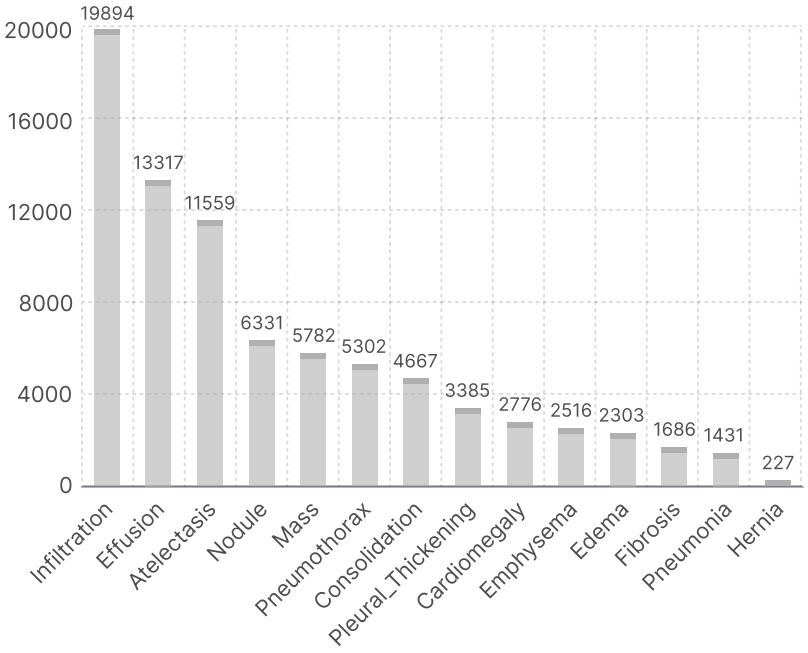


**Figure 9: The label distribution of ChestX-ray14.**

**Table 11: The co-occurrence matrix of ChestX-ray14.**

| | **Atl** | **Car** | **Con** | **Ede** | **Eff** | **Emp** | **Fib** | **Her** | **Inf** | **Mas** | **Nod** | **Ple** | **Pne** | **Ptx** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Atl** | — | 370 | 1,223 | 221 | 3,275 | 424 | 220 | 40 | 3,264 | 739 | 590 | 496 | 262 | 774 |
| **Car** | 370 | — | 169 | 127 | 1,063 | 44 | 52 | 7 | 587 | 102 | 108 | 111 | 41 | 49 |
| **Con** | 1,223 | 169 | — | 162 | 1,287 | 103 | 79 | 4 | 1,221 | 610 | 428 | 251 | 123 | 223 |
| **Ede** | 221 | 127 | 162 | — | 593 | 30 | 9 | 3 | 981 | 129 | 131 | 64 | 340 | 33 |
| **Eff** | 3,275 | 1,063 | 1,287 | 593 | — | 359 | 188 | 21 | 4,000 | 1,254 | 912 | 849 | 269 | 996 |
| **Emp** | 424 | 44 | 103 | 30 | 359 | — | 36 | 4 | 449 | 215 | 115 | 151 | 23 | 747 |
| **Fib** | 220 | 52 | 79 | 9 | 188 | 36 | — | 8 | 345 | 117 | 166 | 176 | 11 | 80 |
| **Her** | 40 | 7 | 4 | 3 | 21 | 4 | 8 | — | 33 | 25 | 10 | 8 | 3 | 9 |
| **Inf** | 3,264 | 587 | 1,221 | 981 | 4,000 | 449 | 345 | 33 | — | 1,159 | 1,546 | 750 | 605 | 946 |
| **Mas** | 739 | 102 | 610 | 129 | 1,254 | 215 | 117 | 25 | 1,159 | — | 906 | 452 | 71 | 431 |
| **Nod** | 590 | 108 | 428 | 131 | 912 | 115 | 166 | 10 | 1,546 | 906 | — | 411 | 70 | 341 |
| **Ple** | 496 | 111 | 251 | 64 | 849 | 151 | 176 | 8 | 750 | 452 | 411 | — | 48 | 289 |
| **Pne** | 262 | 41 | 123 | 340 | 269 | 23 | 11 | 3 | 605 | 71 | 70 | 48 | — | 41 |
| **Ptx** | 774 | 49 | 223 | 33 | 996 | 747 | 80 | 9 | 946 | 431 | 341 | 289 | 41 | — |

※ **Atl** = Atelectasis, **Car** = Cardiomegaly, **Con** = Consolidation, **Ede** = Edema, **Eff** = Effusion, **Emp** = Emphysema, **Fib** = Fibrosis, **Her** = Hernia, **Inf** = Infiltration, **Mas** = Mass, **Nod** = Nodule, **Ple** = Pleural_Thickening, **Pne** = Pneumonia, **Ptx** = Pneumothorax

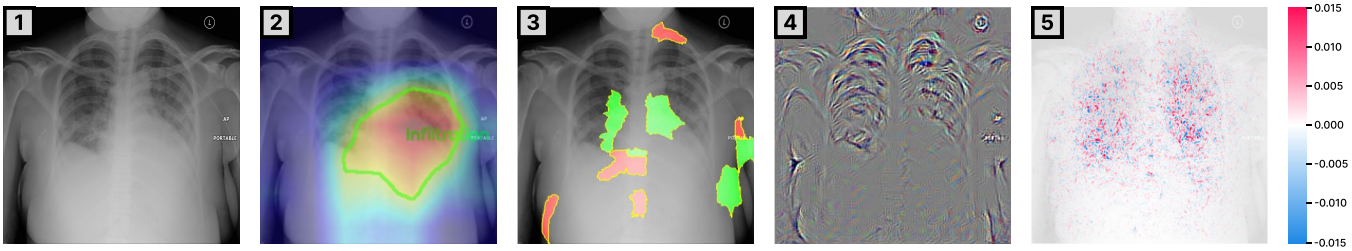


**Figure 10: The comparison of different visualization and interpretability methods is as follows: 1 Original Chest X-ray: Provide the baseline image for analysis. 2 Grad-CAM: Highlights the most influential regions in a coherent, high-level manner. 3 LIME (Top 10 Regions): Delivers fragmented, isolated areas that can be difficult to interpret within a broader diagnostic context. 4 Guided Backpropagation: May introduce noise, potentially obscuring clinically relevant features, particularly in multi-label settings. 5 SHAP: Often results in dispersed or sparse patterns, making interpretation more challenging.**

**Table 12: Evaluation Metrics and Justifications.**

| Evaluation Metrics | Selected | Justification |
|---|---|---|
| Precision (3) | ✓ | Focus on poor performance in underrepresented labels. |
| Recall/Sensitivity (4) | ✓ | Focus on poor performance in underrepresented labels. |
| F1 Score (5) | ✓ | Balanced measure between precision and recall. |
| AUC (6) | ✓ | Measurement on model's discriminatory ability between classes. |
| Accuracy (8) | ✗ | Score inflation tendency due to majority labels. |
| Specificity (9) | ✗ | Score inflation tendency due to majority labels. |
| Mean Average Precision (mAP) (10) | ✗ | Not tailed to one specific label. |

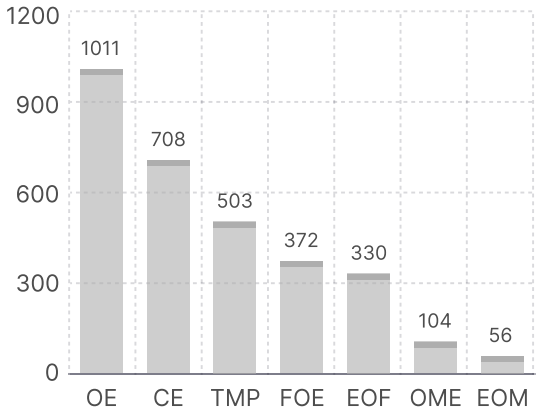


**Figure 11: The label distribution of EarEndo.**

**Table 13: The co-occurrence matrix of EarEndo.**

| | **OME** | **EOF** | **OE** | **EOM** | **FOE** | **CE** | **TMP** |
|---|---|---|---|---|---|---|---|
| **OME** | — | 0 | 10 | 0 | 0 | 0 | 0 |
| **EOF** | 0 | — | 41 | 0 | 0 | 26 | 8 |
| **OE** | 10 | 41 | — | 5 | 9 | 28 | 8 |
| **EOM** | 0 | 0 | 5 | — | 0 | 6 | 0 |
| **FOE** | 0 | 0 | 9 | 0 | — | 75 | 0 |
| **CE** | 0 | 26 | 28 | 6 | 75 | — | 5 |
| **TMP** | 0 | 8 | 8 | 0 | 0 | 5 | — |

※ **OME** = Otitis Media with Effusion, **EOF** = External Auditory Canal Foreign Body, **OE** = Otitis Externa, **EOM** = External Auditory Canal Tumor, **FOE** = Fungal Otitis Externa, **CE** = Cerumen Impaction, **TMP** = Tympanic Membrane Perforation