Towards Multi-Label Concept Bottleneck Models in Medical Imaging: An Exploratory Survey

Berthine Nyunga Mpinda[1] BERTHINE.NYUNGA-MPINDA@UNI-TUEBINGEN.DE
Mehran Hosseinzadeh[1] MEHRAN.HOSSEINZADEH@UNI-TUEBINGEN.DE
Valay Bundele[1] VALAY.BUNDELE@UNI-TUEBINGEN.DE
Hendrik P. A. Lensch[1] HENDRIK.LENSCH@UNI-TUEBINGEN.DE
[1] *University of Tuebingen*

Editors: Under Review for MIDL 2026

Abstract

Deep neural networks achieve strong performance in medical image classification, but their lack of interpretability hinders clinical adoption. Concept Bottleneck Models (CBMs) address this by predicting human-understandable intermediate concepts, yet prior work has focused almost exclusively on single-label tasks and has not examined CBMs under the multi-label conditions typical of medical imaging. Because multiple concepts may appear in an image regardless of the final task, CBMs are highly sensitive to class imbalance, co-occurring pathologies, and concept noise. We present the first systematic study of label-free CBMs for multi-label chest X-ray classification, using LLMs to generate concepts and VLMs to align them with images. We evaluate performance, robustness, and interpretability under realistic clinical conditions, analyzing long-tail label distributions, label co-occurrence, and the impact of VLM choice on concept alignment and downstream prediction. Experiments show that label-free CBMs achieve competitive AUROC but reduced precision on minority classes, with performance strongly influenced by backbone selection and class imbalance. Medical-domain VLMs (e.g., BiomedCLIP, BioViL, RAD-DINO) provide consistent gains, and balanced losses improve minority-class sensitivity. Concept-level analysis indicates that pseudo-concepts are reliable for common conditions but unstable for rare ones, affecting multi-label prediction quality. Overall, this work offers the first comprehensive evaluation of label-free CBMs in multi-label medical imaging and practical guidelines for building interpretable, clinically meaningful models.

Keywords: Concept bottleneck models, Multi-label classification, vision-language models, interpretability.

1. Introduction

Artificial intelligence (AI) has significantly advanced medical image analysis, improving diagnostic speed, accuracy, and consistency across modalities such as chest X-ray and CT (Faiyazuddin et al., 2025). Despite strong performance, deep neural networks (DNNs) remain "black boxes," limiting clinical trust and adoption in safety-critical settings (Kim et al., 2023b; Yan et al., 2023b; Chowdhury et al., 2024). Concept Bottleneck Models (CBMs) (Koh et al., 2020) offer an interpretable alternative by predicting human-understandable concepts before making a final decision. However, traditional CBMs require expert-annotated concepts, which are costly and difficult to scale in medical imaging. Label-free CBMs address this by generating concepts automatically using large language models

(LLMs) and aligning them with images using vision–language models (VLMs) (Oikarinen et al., 2023; Yang et al., 2023).

Despite this progress, current CBM research has notable limitations. (1) Prior work focuses almost exclusively on single-label classification for natural images or isolated medical tasks (Yang et al., 2023; Oikarinen et al., 2023; Yan et al., 2023b), leaving multi-label medical imaging largely unexplored. (2) The influence of different VLMs—such as CLIP, BiomedCLIP, BioViL, and RAD-DINO—on concept alignment and downstream CBM behavior remains poorly understood. (3) Existing evaluations rarely incorporate multi-label metrics or account for severe class imbalance, making it unclear whether current CBM findings extend to realistic clinical scenarios.

Importantly, CBMs are inherently multi-label systems, because multiple concepts may be simultaneously present in an image—even when the downstream task is single-label. In medical imaging, where pathologies frequently co-occur, this CBM-induced multi-labeledness interacts with real-world factors such as long-tail distributions, correlated abnormalities, and concept noise. This makes multi-label evaluation not only relevant but essential. In this work, we conduct the first systematic study of label-free CBMs for multi-label chest X-ray classification, exploring how pseudo-concepts behave under real-world conditions. Using the NIH ChestX-ray14 dataset (Wang et al., 2017), we examine CBM performance, imbalance robustness, and concept reliability across several medical-domain VLMs.

Specifically, we investigate: How well do label-free CBMs perform on multi-label medical image classification? How does CBM performance change under extreme class imbalance and long-tail prevalence? How reliable and clinically meaningful are the generated (pseudo-) concepts? How strongly does VLM choice influence concept alignment and downstream predictions?

This work makes the following contributions:

- We present the first comprehensive evaluation of label-free CBMs on a multi-label medical imaging task.
- We analyze how class imbalance, label co-occurrence, and long-tail distributions affect CBM performance and interperability, revealing that the often-used AUC metric is a poor indicator of the multi-label prediction quality.
- We compare several medical-domain VLMs—including BiomedCLIP, BioViL, BioViL-T, and RAD-DINO—and show their impact on multi-label CBM behavior.

Overall, this study provides the first thorough understanding of how label-free CBMs operate under realistic multi-label clinical conditions and outlines practical directions for building more reliable, interpretable, and clinically meaningful CBM-based systems.

2. Related work

2.1. Multi-Label Classification

Multi-label learning addresses problems where each instance can be associated with multiple labels simultaneously, requiring models to account for correlations among labels rather than assigning a single class (Ju et al., 2024; Wu et al., 2020; Huang et al., 2021; Zhang and Wu,

2024). This makes multi-label classification (MLC) inherently more challenging than binary or multi-class tasks. A major difficulty in MLC is the long-tailed distribution of labels: a few classes dominate the dataset, while many others are rare (Ju et al., 2023; Holste et al., 2023; Cao et al., 2019). This imbalance is especially pronounced in medical imaging, where some diseases are rare and collecting large numbers of examples is difficult (Ju et al., 2023, 2024). For example, chest radiographs often contain multiple abnormalities, yet most images reflect only a few common conditions (Holste et al., 2024; Ridnik et al., 2021; Irtaza et al., 2024).

MLC has been widely explored in image, text, and biomedical domains (Li et al., 2017). Classical approaches include binary relevance (BR) (Zhang et al., 2018), which treats labels independently, and methods that model label dependencies, such as classifier chains (Read et al., 2011) and RNN-based label modeling (Yao et al., 2017). To address class imbalance, several strategies have been proposed. **Data-based methods** such as over- and under-sampling can help but risk overfitting rare classes or increasing noise (Wu et al., 2020; Fang et al., 2023). **Loss-based methods** modify the objective to emphasize minority labels, including Focal Loss (Dong, 2020), Distribution-Balanced Loss (Wu et al., 2020), asymmetric and two-way losses (Ridnik et al., 2021; Holste et al., 2024), contrastive-based formulations (Zhang and Wu, 2024), hierarchical penalty losses (Asadi et al., 2025), and rank-based losses targeting rare labels (Hanif et al., 2025). **Modeling-based approaches** such as bag of multi-label descriptors (BoMD) (Chen et al., 2023) and multi-label boosting (Thach et al., 2025) aim to better capture label co-occurrence and improve tail-class performance. Despite extensive work, no single method fully resolves the combined challenges of label correlation, rarity, and multi-pathology co-occurrence, particularly in medical data. These limitations have motivated our evaluation of loss-based strategies within a label-free CBM setting to better understand their effectiveness in multi-label medical imaging.

2.2. Multi-Label Classification in Medical Imaging

Multi-label classification is essential in medical imaging, where a single sample often contains co-occurring abnormalities. Chest X-rays exemplify this setting, with common combinations such as effusion, consolidation, and edema. This introduces three key challenges: (i) severe long-tailed distributions, with a few common diseases dominating and many clinically important ones remaining rare (Wang et al., 2017; Holste et al., 2024; Irtaza et al., 2024); (ii) strong label dependencies that reflect meaningful clinical co-occurrence patterns (Asadi et al., 2025; Kumar et al., 2018); and (iii) label noise and uncertainty, particularly in weakly annotated datasets such as NIH ChestX-ray14 (Cid et al., 2024; Efimovich et al., 2024).

To address these difficulties, recent work has explored a range of architectures and training strategies. CNN-based and transfer learning models—including ResNet, EfficientNet, DenseNet, and ConvNeXt—remain widely used for multi-label prediction (Pillai, 2022; Kufel et al., 2023; Khan et al., 2024). Competitive results in recent benchmarks, such as the CXR-LT Challenge (Holste et al., 2024), have relied on ImageNet-pretrained backbones such as ConvNeXt, EfficientNetV2, ResNeXt101, combined with standard techniques such as augmentation, class weighting, and model ensembling. A variety of methods have been proposed to mitigate class imbalance and capture label relationships. Loss-based strategies—such as focal loss, distribution-balanced loss, asymmetric variants, and ranking-based objectives like FZLPR—aim to improve performance on rare classes (Wu et al., 2020; Hanif

et al., 2025). Other approaches incorporate hierarchical or cascaded structures to explicitly model label dependencies (Asadi et al., 2025; Kumar et al., 2018). Multimodal solutions further reduce label noise by aligning visual features with text reports, including frameworks such as X-RayDar (Cid et al., 2024) and vision–language models like BiomedCLIP-PubMedBERT (Ganapathy et al., 2024).

Despite this progress, multi-label medical imaging remains challenging due to rare disease classification, correlated abnormalities, and noisy or incomplete labels. These issues are particularly impactful for concept-based models, where intermediate concept quality directly shapes final predictions. In label-free CBMs, concepts generated via large language models and aligned through vision–language models add an additional layer of sensitivity to imbalance and noise. Motivated by these challenges, we evaluate four imbalance-aware loss functions from the multi-label literature to assess their effect on CBM performance in a medical multi-label task.

2.3. Vision-Language Models (VLMs)

Foundation models—large neural networks pretrained on massive multimodal datasets—have transformed machine learning by enabling broad generalization and flexible adaptation across downstream tasks (Minaee et al., 2024). VLMs extend this capability by aligning visual and textual representations, making them especially useful for interpretable visual reasoning and zero-shot recognition. CLIP (Radford et al., 2021), a prototypical VLM, uses paired image–text encoders to project both modalities into a shared semantic space, enabling strong generalization and serving as the basis for many adaptation methods such as prompt learning (Khattak et al., 2025).

In the medical domain, general-purpose VLMs often struggle because clinical images exhibit subtle pathologies that differ from natural images. To address this mismatch, several domain-adapted VLMs have emerged. BiomedCLIP (Zhang et al., 2023) leverages large-scale biomedical image–text pairs to improve semantic grounding in clinical terminology, while BioViL (Boecking et al., 2022) and BioViL-T (Bannur et al., 2023) incorporate radiology reports during pretraining to strengthen alignment between chest X-rays and radiological concepts. These medical VLMs have demonstrated superior performance in tasks such as zero-shot classification, report retrieval, and disease localization, making them particularly suitable for concept-based modeling in medical imaging.

2.4. Concept Bottleneck Models (CBMs)

Concept Bottleneck Models (CBMs) (Koh et al., 2020) replace black-box prediction with a two-step process: first predicting human-interpretable concepts, then using them to infer the final label. This modular structure offers transparency, supports test-time interventions, and improves error analysis, making CBMs particularly attractive in high-stakes domains such as medical imaging (Oikarinen et al., 2023; Yang et al., 2023). CBMs require tuples (x, c, y) during training, where a concept encoder predicts $\hat{c} = g(x)$ and a label predictor maps these concepts to $\hat{y} = f(\hat{c})$. Different training regimes (sequential, independent, joint) achieve comparable accuracy (Koh et al., 2020), though independently trained CBMs better support human correction of predicted concepts (Gupta and Narayanan, 2024).

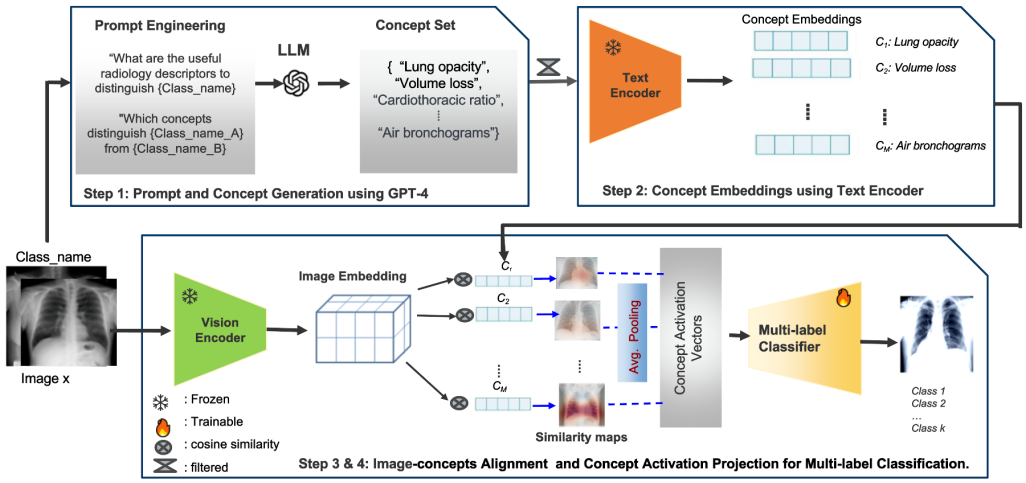

Figure 1: Pipeline Overview for Multi-Label Label-Free CBM Training.

CBMs can be trained in concept-supervised or label-free settings. Concept-supervised CBMs rely on predefined, annotated concepts and provide strong interpretability and controllability. Extensions include probabilistic CBMs (Kim et al., 2023a), graph-structured concept reasoning (Xu et al., 2025), and prototype-based visual grounding for spatial interpretability (Huy et al., 2025). Semi-supervised variants further reduce annotation demands by aligning unlabeled data with concept predictions (Hu et al., 2025).

Label-free CBMs remove the need for manual concept annotation by generating concepts automatically—typically with LLMs—and aligning them with images using vision–language models (Oikarinen et al., 2023). Later work improves concept selection and noise reduction through submodular optimization (Yang et al., 2023), visual activation scoring (Kim et al., 2023b), and object-centric grounding (Steinmann et al., 2025). These methods achieve performance competitive with supervised CBMs while remaining scalable.

CBMs have been applied to natural image datasets such as CUB, CIFAR, and ImageNet, and increasingly to medical imaging tasks including chest X-rays, histopathology, and dermatology (Sun et al., 2025; Yan et al., 2023b). While concept-supervised CBMs offer precise, faithful explanations, label-free CBMs provide broader scalability. Key remaining challenges include reducing concept noise, ensuring faithful reasoning, and extending CBMs to complex multi-label medical settings—motivating the evaluation conducted in this work.

3. Methodology

In this section, we present an automated pipeline for training a label-free CBM for multi-label chest X-ray classification. The approach integrates four components: (i) clinically meaningful concepts automatically generated using an LLM and filtered, (ii) concept embeddings obtained via a pretrained text encoder, (iii) concept–image alignment scores computed by pretrained VLMs, and (iv) a downstream multi-label classifier operating on the resulting concept bottleneck representation. The complete pipeline is illustrated in Figure 1.

3.1. Problem Setting

Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times C}$ is an input image of height H, width W, and C channels, and $y_i \in \{0,1\}^M$ is a binary vector indicating the presence or absence of M possible labels. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_T\}$ denote the set of visual concepts obtained from LLMs or expert-defined sources. We use a vision encoder $E_I : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$ and a text encoder $E_T : \mathcal{C} \to \mathbb{R}^{T\times d}$, both initialized with pretrained parameters from large-scale image or image-text datasets. The concept scores S_i for an image x_i are computed as:

$$S_i = E_I(x_i) \cdot E_T(\mathcal{C})^\top. \tag{1}$$

The concept score $S_i \in \mathbb{R}^T$ represents the similarity between the image and each concept, and $\cdot$ denotes the dot product or cosine similarity. Then, an interpretable multi-label classifier $f : S = S_i \to Y$ predicts labels based on the concept activations.

3.2. Concept Generation via LLMs

Given a class label y_i, we query a large language model (GPT-4) to generate a candidate set of concepts $\mathcal{C}_i^{\text{raw}}$. Following prior work (Oikarinen et al., 2023; Yan et al., 2023b), the queries were formulated as:

```
"What are the useful radiology descriptors to distinguish {class_name}?"
"Which concepts distinguish {class_name_A} from {class_name_B}?"
```

where `class_name` refers to pathologies such as Atelectasis or Pneumonia. To ensure clinical relevance, the raw concepts are filtered automatically by: (i) removing excessively long phrases, (ii) eliminating redundant or overlapping entries, and (iii) discarding concepts too similar to the original pathology name. The filtered set $\mathcal{C}_j^{\text{raw}}$ is then used as $\mathcal{C}_i = \text{Filter}(\mathcal{C}_i^{\text{raw}})$.

After filtering, the final concept set contains 56 distinct concepts. For example, a list of concepts generated for the pathology "Mass" and "Cardiomegaly" includes:

```
Mass:
- "pulmonary mass",
- "space occupying lesion",
- "well defined opacity",
- "lobulated density".
```

```
Cardiomegaly:
 - "enlarged cardiac silhouette",
 - "increased cardiothoracic ratio",
 - "prominent heart size".
```

These concepts are subsequently encoded via a pretrained text encoder E_T to obtain embeddings. The resulting concept embeddings serve as the intermediate, interpretable bottleneck for our label-free CBM, enabling multi-label classification without requiring manual concept annotations.

3.3. Concept-Image Alignment

To align the image features with the generated concepts, we computed the cosine similarity between their embeddings.

$$S_{i,j} = \frac{E_I(x_i) \cdot E_T(c_j)}{\|E_I(x_i)\| \, \|E_T(c_j)\|}, \quad S_i = [S_{i,1}, S_{i,2}, \ldots, S_{i,|\mathcal{C}_i|}], \tag{2}$$

where $E_I(x_i) \in \mathbb{R}^d$ is the image embedding, $E_T(c_j) \in \mathbb{R}^d$ is the concept embeddings, $\cdot$ denotes the dot product, and $\|\cdot\|$ is the ℓ_2 norm. S_i contains the similarity scores between the image embedding and each concept embedding. Higher values indicate a strong correlation. We assume E_I and E_T are initialized with pretrained parameters from large-scale image or image-text datasets, such as ImageNet or medical imaging datasets, providing strong multi-modal embeddings for computing meaningful concept scores.

We selected a variety of backbones and VLMs such as ResNet-50 (baseline) and CLIP (Radford et al., 2021), trained on ImageNet, as well as specialized VLMs for the medical domain, including BioViL (Boecking et al., 2022), BioViL-T (Bannur et al., 2023), and BioMedCLIP (Zhang et al., 2023). Additionally, we use RAD-DINO (Perez-Garcia et al., 2025), a vision transformer trained to encode chest X-rays via the self-supervised method DINOv2 (Oquab et al., 2023). These VLMs were selected based on their reported performance in chest X-ray classification and segmentation tasks (Yan et al., 2023b; Ganapathy et al., 2024; Srivastava et al., 2024; Li et al., 2025) and because several have been explicitly tuned for radiology imaging. Table 3 shows more details about their image and text encoders. Since RAD-DINO is an image-only encoder, we adopt the approach from (Barsellotti et al., 2025) and use the BioMedCLIP text encoder to obtain concept embeddings. These embeddings are then aligned with RAD-DINO's patch-level image features via similarity computation, resulting in concept activation vectors.

3.4. Multi-Label Classification

Finally, we train a classifier f_θ on the fixed concept scores S_i to predict the multi-label vector $\hat{y}_i$:

$$\hat{y}_i = f_\theta(S_i), \quad \hat{y}_i \in [0,1]^M. \tag{3}$$

where a sigmoid activation is applied to each output dimension to handle multi-label predictions independently. The model is trained using the binary cross-entropy (BCE) loss over all labels:

$$\mathcal{L}_{\text{BCE}}(y_i, \hat{y}_i) = -\frac{1}{M} \sum_{j=1}^{M} \Big[y_{ij} \cdot \log \hat{y}_{ij} + (1 - y_{ij}) \cdot \log(1 - \hat{y}_{ij}) \Big]. \tag{4}$$

The CBM uses the VLM-derived concept scores S_i as a fixed, interpretable bottleneck representation. Since concepts are inferred directly through VLM image–text similarity, the concept predictor contains no trainable parameters. This design fully decouples concept extraction from classifier training and guarantees that all downstream predictions must pass through the concept space, thereby preserving interpretability as shown in Steps 3 and 4 of the pipeline in Figure 1.

For multi-label classification, we train a multilayer perceptron (MLP) that maps the concept vector to the 14 disease labels. The MLP takes $S_i \in \mathcal{R}^k$ as input and outputs a probability vector in $[0,1]^{14}$ using sigmoid activations to predict the final pathologies.

To address label imbalance, we use loss functions that take into account the long-tailed distribution such as weighted binary cross-entropy (BCE), focal loss (Ross and Dollár, 2017),

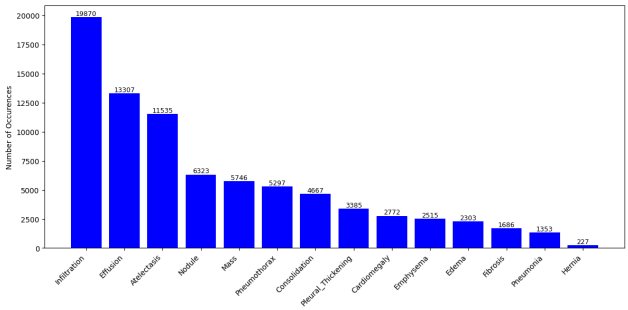

Figure 2: Label distribution of pathologies in the NHI ChestX-ray dataset.

and two-way multi-label loss (Kobayashi, 2023):

$$\mathcal{L}_{\text{WBCE}}(y_i, \hat{y}_i) = -\frac{1}{M}\sum_{j=1}^{M}\Big[w_j \cdot y_{ij} \cdot \log \hat{y}_{ij} + (1 - y_{ij}) \cdot \log(1 - \hat{y}_{ij})\Big], \tag{5}$$

$$\mathcal{L}_{\text{Focal}}(y_i, \hat{y}_i) = \alpha \cdot (1 - p_t)^{\gamma} \cdot \mathcal{L}_{\text{BCE}}(y_i, \hat{y}_i), \tag{6}$$

$$\mathcal{L}_{\text{Two-way}}(y_i, \hat{y}_i) = \frac{1}{M}\sum_{i=1}^{M} l_i \cdot (\{x_{ic}, y_{ic}\}_{c=1}^{C}; T) + \frac{1}{C}\sum_{c=1}^{C} l^c \cdot (\{x_{ic}, y_{ic}\}_{i=1}^{M}; T), \tag{7}$$

where w_j are the class weights, α balances the classes, p_t is the predicted probability for the true label, $\gamma \geq 0$ is the focusing parameter, and T is the temperature. l_i and l^c denote the sample-wise and class-wise components of the two-way loss, respectively.

4. Experiments and Results

4.1. Dataset

To evaluate and validate multi-label CBM models, we conduct experiments on the NIH ChestX-ray14 dataset (Wang et al., 2017), one of the largest publicly available chest X-ray collections. The dataset contains 112,120 frontal-view radiographs from 30,805 unique patients collected between 1992 and 2015 at the National Institutes of Health Clinical Center, MD, USA. Each image is annotated with up to 14 thoracic pathologies, extracted from associated radiological reports using natural language processing (NLP) techniques (Wang et al., 2017). The 14 thoracic pathologies are: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. The dataset exhibits a highly imbalanced label distribution, reflecting the unequal prevalence of different conditions. For example, Hernia and Pneumonia appear in only 0.2% and 1.2% of the images, respectively as shown in Figure 2. We use the official patient-wise split provided by (Wang et al., 2017): 70% training, 10% validation, and 20% testing. This split has been widely used in previous work for classical multi-label classification (Huang et al., 2024; Yao et al., 2017; Rajpurkar, 2017; Guendel et al., 2019; Yao et al., 2018).

4.2. Training Procedure

We train only the label-prediction multi-layer perceptron, as the concept activation vectors are static. The model is initially optimized by binary cross-entropy (BCE), as commonly used for multi-label classification. Given the strong class imbalance in ChestX-ray14, several strategies are applied to stabilize training and improve calibration, such as data augmentation and optimal thresholding per label. To mitigate class imbalance, we apply losses such as WBCE loss, focal loss and two-way multi-label loss. We employ Adam optimizer with a learning rate of $1 \times e - 3$ and a batch size of 256 for 100 epochs. To prevent overfitting, we incorporate dropout, weight decay, and early-stopping based on validation loss.

4.3. Evaluation Metrics

We evaluate model performance using a set of complementary metrics suited for multi-label medical image classification. Although AUC-ROC is traditionally the dominant metric in chest X-ray benchmarks (Wang et al., 2017; Yao et al., 2017; Hanif et al., 2025), it can be overly optimistic for unbalanced datasets such as ChestX-ray14. Therefore, we report AUC along with threshold-dependent metrics, i.e. precision, recall, F1 score, and per-class confusion matrices to better reflect clinically meaningful behavior.

AUC and ROC Analysis: For each disease label, we compute the ROC curve and its Area Under the Curve (AUC), which measures the model's ability to rank positive samples higher than negative ones across all thresholds. We report per-class AUC for each pathology, macro AUC, averaging all classes equally and micro AUC, pooling predictions across labels, which is dominated by majority diseases. Although AUC is threshold-independent, it can mask poor precision or recall for rare classes, motivating the use of additional metrics.

Precision, Recall, and F1 Score: To assess threshold-sensitive performance, which is particularly important in imbalanced settings, we report precision, recall, and the F1 score for each label. We compute the macro average for F1, precision and recall, across diseases in order to reflect the performance on rare diseases. Similarly, we compute the micro F1, precision and recall, aggregated across all predictions. These metrics highlight trade-offs between false positives and false negatives that AUC alone cannot capture. They also reveal the impact of dominating diseases in the final prediction.

Confusion Matrix Analysis: We analyze the per-class confusion matrices, summarizing the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each pathology. This analysis allows us to identify systematic underdetection of rare diseases (high false negatives), false positives driven by visually similar conditions, and class-specific precision–recall trade-offs. Such error-level insights are essential for understanding the failure modes of multi-label classification in clinical decision-making.

4.4. Results and Analysis

4.4.1. Multi-Label Classification Performance comparison

We conduct a comprehensive comparison of multi-label classification performance using the selected VLMs, including ResNet50, CLIP, BioMedCLIP, BioViL, BioViL-T, and RAD-DINO, for image and concept embedding extraction. Performance is evaluated using eight metrics: macro/micro AUC, macro/micro F1, macro/micro precision, and macro/micro

Table 1: Comparisons of different backbones in CBMs. The best scores are highlighted in bold and the second-best is underlined.

Backbone	AUC-mi	AUC-ma	F1-mi	F1-ma	Pr-mi	Pr-ma	Re-mi	Re-ma
ResNet50	0.7304	0.5067	0.3556	0.1105	0.2564	0.0869	0.5801	0.2543
CLIP	0.7836	0.6256	0.3125	0.2210	0.2005	0.1462	**0.7074**	**0.5353**
BiomedCLIP	<u>0.8296</u>	<u>0.7375</u>	<u>0.4163</u>	<u>0.3014</u>	<u>0.3178</u>	<u>0.2437</u>	0.6031	0.4483
BioViL	0.7883	0.6436	0.3430	0.2131	0.2345	0.1513	<u>0.6384</u>	0.4120
BioViL-T	0.7838	0.6297	0.3430	0.2041	0.2382	0.1651	0.6127	0.3762
RAD-DINO	**0.8823**	**0.8220**	**0.5054**	**0.4076**	**0.4345**	**0.3643**	0.6079	<u>0.4908</u>

Table 2: Comparisons of CBMs with different embedding types (global, patch, combined).

Backbone	Type	AUC-mi	AUC-ma	F1-mi	F1-ma	Pr-mi	Pr-ma	Re-mi	Re-ma
BioViL	Global	**0.7883**	**0.6436**	<u>0.3430</u>	**0.2131**	0.2345	<u>0.1513</u>	0.6384	<u>0.4120</u>
	Patch	0.7743	0.6045	0.3171	0.2003	0.2115	0.1399	0.6326	0.4092
	Comb.	0.7828	0.6277	0.3337	<u>0.2082</u>	0.2280	0.159	0.6222	0.4075
ResNet50	Global	0.7304	0.5067	**0.3556**	0.1105	<u>0.2564</u>	0.0869	0.5801	0.2543
	Patch	0.7263	0.5000	0.2971	0.1405	0.1890	0.0883	**0.6945**	0.4029
	Comb.	0.7378	0.5017	0.3371	0.0996	**0.2648**	0.0760	0.4638	0.1959
BioViL-T	Global	<u>0.7838</u>	<u>0.6297</u>	0.3430	0.2041	0.2382	**0.1651**	0.6127	0.3762
	Patch	0.7783	0.6115	0.3274	0.2071	0.2159	0.1414	<u>0.6768</u>	**0.4350**
	Comb.	0.7820	0.6196	0.3401	0.2076	0.2363	0.1476	0.6064	0.3751

recall. To ensure a fair comparison, all models are trained under the same settings, and all image and text encoders used for feature extraction are frozen.

Table 1 summarizes performance on the test set. Across all backbones, AUC values are high, consistent with previous literature on ChestX-ray14. In contrast, metrics such as F1, precision, and recall reveal larger differences between VLMs, reflecting the challenges posed by class imbalance in test data. Domain-specific VLMs, including BioMedCLIP, BioViL, BioViL-T, and RAD-DINO, outperform ResNet50 and CLIP, indicating that medical-domain alignment improves concept-based prediction quality. Among all backbones, RAD-DINO achieves the highest overall performance, obtaining the best macro and micro AUC of of 0.8823% and 0.8220%, respectively, as well as the highest F1 and precision in most cases. This demonstrates that combining a self-supervised, radiology-specific image encoder with domain-tuned concept embeddings can significantly enhance multi-label classification in a label-free CBM framework. Overall, these results suggest that the choice of backbone has a critical impact on both ranking-based and threshold-dependent performance, and that domain-specific backbones are particularly effective in capturing clinically relevant patterns in chest X-ray images.

Performances with three embedding strategies (global, patch, combined) in Table 2 show that global embeddings consistently emerged as the best strategy. Global embeddings provide the most reliable overall performance, patch embeddings improved recall at the cost of precision, and combined embeddings offer no consistent advantage. Medical-domain VLMs (BioViL, BioViL-T) consistently outperform ResNet50, particularly in macro-AUC and macro-F1, indicating better handling of rare pathologies. The best results were achieved by BioViL with global embeddings (micro/macro AUC: 0.7883/0.6436) and BioViL-T with patch embeddings (macro-recall: 0.435). These results indicate that medical VLM back-

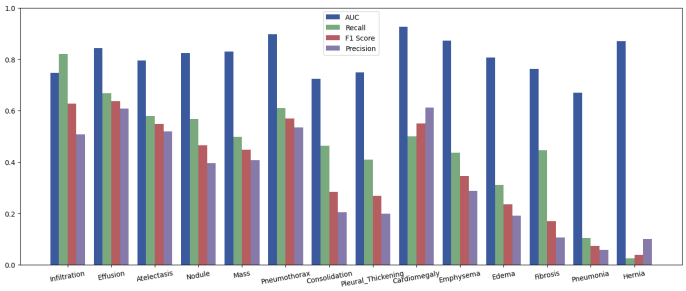

Figure 3: AUC, F1, Recall and Precision performance for each pathology. Note the strong discrepancy between the almost constant AUC vs. the strong drop in the other metrics for less frequent classes.

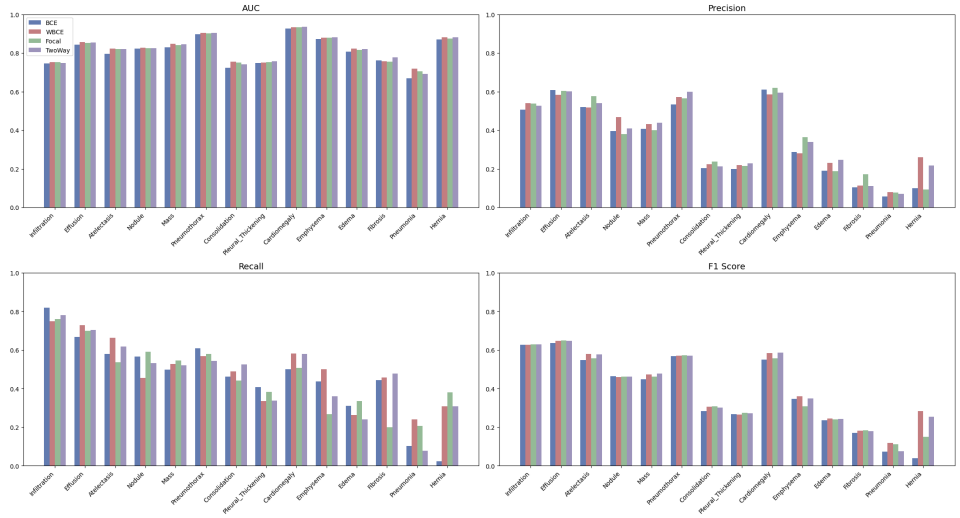

Figure 4: Performance improvement across different loss functions.

bones combined with global embeddings provide the strongest and most stable CBM performance.

4.4.2. Per-Label Performance Analysis

To further investigate model behavior, we evaluated per-label metrics for all 14 thoracic pathologies. Figure 5 shows the heatmap of AUC and precision for each label using the six VLMs. We observe that the AUC remains relatively high across most classes while precision scores vary considerably, highlighting the impact of class imbalance. Rare diseases such as Hernia and Pneumonia show lower precision despite competitive AUC, indicating that ranking performance is good, but absolute positive prediction remains challenging. This per-label analysis provides insight into specific failure modes and helps identify pathologies that may benefit from additional concept refinement or threshold tuning.

Zooming in on performance metrics for some majority classes such as Infiltration, Atelectasis, and Effusion, and minority classes such as Pneumonia, Fibrosis, and Hernia, which

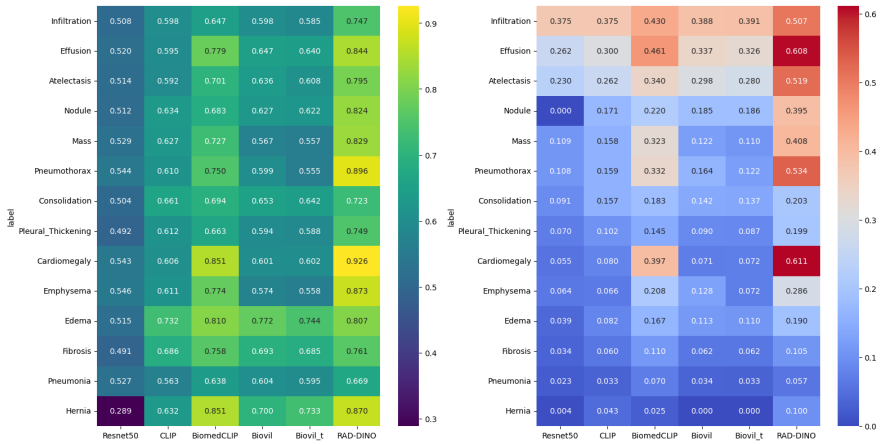

Figure 5: Per-label AUC scores and precision for each thoracic pathology with different VLMs. High AUC scores are observed even for rare diseases, while their precisions are poor, especially the minority classes such as Pneumonia and Hernia.

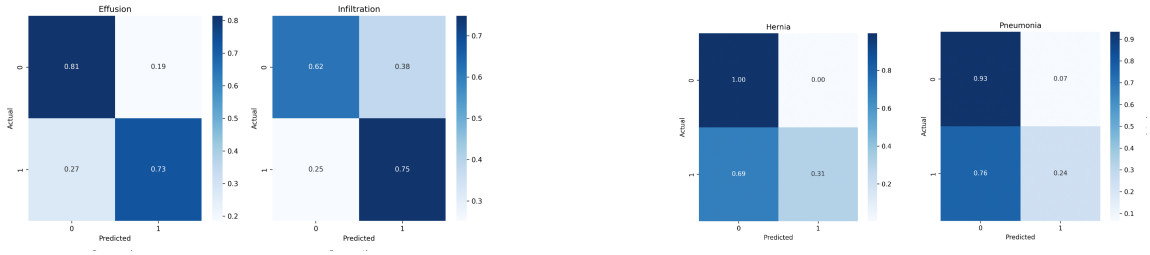

Figure 6: Confusion matrices of some majority classes (Effusion and Infiltration) and minority classes (Pneumonia and Hernia).

together represent less than 5% of the dataset, Figure 3 highlights a substantial gap between AUC and other metrics. For the majority classes, AUC is high and precision, recall and F1 are relatively strong (around 50%), indicating that the model ranks positives above negatives effectively and can reliably identify common pathologies. Additionally, from the confusion matrices in Figure 6, 70% of images are correctly predicted. In contrast, for minority classes, while AUC remains high, precision, recall, and F1 are much lower.

This shows that the model can rank rare positive instances correctly (hence high AUC) but struggles to generate accurate predictions at practical thresholds, often producing false negatives and false positives. The confusion matrices in Figure 6 show how the model produces a high percentage of false negative for the minority classes Pneumonia and Hernia.

4.4.3. Fighting the Effects of Class Imbalance

In the medical domain, the performance gap shown in Figures 3 and 6 is critical. Missing rare but clinically significant pathologies leads to undiagnosed conditions, while too many false positives may overburden radiologists with recurrent follow-ups. Thus, strategies such as per-label threshold tuning, data augmentation, and class-balanced losses such as WBCE (Equation 5), focal loss (Equation 6), and two-way multi-label loss (Equation 7) are applied.

Figure 4 illustrates the impact of these losses. While improvements in AUC were modest, precision, recall, and F1 scores showed more substantial gains. For minority classes such as Pneumonia and Hernia, these gains are especially meaningful, as they increase the likelihood of correctly detecting rare diseases. In particular, WBCE and two-way loss effectively emphasize minority classes, helping the model better handle class imbalance in high-stakes medical tasks. Addressing class imbalance is critical when training CBMs for multi-label medical prediction. We recommend using class-balanced loss functions, such as weighted BCE or two-way loss, to ensure that rare pathologies are accurately predicted. Per-label threshold tuning and data augmentation for minority classes can further improve performance while controlling false positives. Evaluations should emphasize minority-class metrics like precision, recall, and F1, alongside overall AUC, to ensure that CBMs reliably detect rare diseases and support high-stakes clinical decision-making.

4.4.4. Interpretability and Concept Quality

CBMs are inherently interpretable, as the learned weights represent the contribution of each concept to a given class prediction. Using our best multi-label classifier, we visualize concept contributions for each label. The horizontal bars in Figure 7 show the top concepts for Infiltration (a majority class) and Hernia (a minority class), highlighting which concepts drive predictions positively or negatively. Figure 8 (in Appendix) presents concept contributions across all labels, providing a compact view of key contributors for the predictions.

Examining concept quality, the top three predicted concepts for Cardiomegaly are correct, while all four top concepts for Infiltration appear within the top-10 concepts. In contrast, for Hernia, none of its concepts appear in the top-10, likely reflecting the low number of samples for this class. Predicting concepts is itself a multi-label problem, and using them to predict multi-label classes remains challenging. Furthermore, as input images can contain multiple labels, the alignment between images and concepts may capture overlapping or co-occurring pathologies, adding complexity to the prediction task.

5. Discussion and Future Work

Multi-label medical CBMs demonstrate clear performance gains when built on top of domain-specialized vision-language models. In our evaluation, specialized VLMs such as BiomedCLIP, RAD-DIO, BioViL, and BioViL-T consistently outperformed the generic ResNet50 backbone, highlighting the importance of representations grounded in medical domains. Among the embedding strategies, global embeddings proved particularly effective, offering stable and coherent representations that better support multi-label prediction. Handling class imbalance also emerged as a critical factor: weighted BCE and two-way

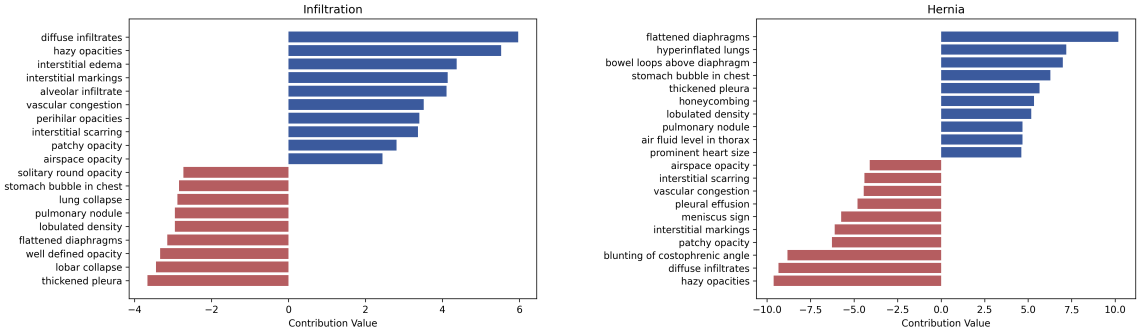

Figure 7: Diverging concept contributions for Infiltration and Hernia, where blue and red bars indicate the top positive and negative contributions, respectively.

loss notably improved minority-class precision, recall, and F1, reducing false positives in rare pathologies. At the interpretability level, concept accuracy played a central role. Rare classes frequently exhibited incomplete or misaligned concepts, which weakened downstream prediction reliability in cases with overlapping abnormalities. Enhancing concept coverage through targeted augmentation, improved sampling, or synthetic data can help stabilize predictions and strengthen the interpretability that CBMs promise.

Looking forward, several avenues could further advance multi-label CBMs in medical imaging. Hybrid concept modeling strategies, where human-supervised concepts are reserved for clinically critical findings while automated or LLM-generated concepts support broader coverage, could balance scalability with clinical fidelity. Semi-supervised learning may help leverage unlabeled data to refine concept bottlenecks, while hierarchical concept structures could enable models to reason across coarse- and fine-grained clinical attributes. Integrating causal concept modeling also holds promise for improving robustness by encouraging models to rely on medically meaningful features rather than dataset-specific correlations. Finally, clinician-in-the-loop validation and interactive concept correction tools could facilitate real-world integration, allowing experts to refine noisy concepts and enhance the transparency and trustworthiness of CBMs in clinical workflows.

6. Conclusion

This work provides the first systematic evaluation of label-free Concept Bottleneck Models for multi-label medical imaging, showing that medical-domain VLMs paired with global embeddings yield the most reliable performance. Although imbalance-aware losses partially recover minority-class precision and recall, concept fidelity, especially for rare diseases, remains a major challenge. Overall, while label-free CBMs can achieve competitive accuracy and interpretable reasoning, advancing concept alignment, improving robustness to long-tail distributions, and incorporating clinically meaningful supervision are essential next steps toward safe and effective deployment in medical practice.

References

Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pages 201–209. Springer, 2025.

Mehrdad Asadi, Komi Sodoké, Ian J Gerard, and Marta Kersten-Oertel. Clinically-inspired hierarchical multi-label classification of chest x-rays with a penalty-based loss function. *arXiv preprint arXiv:2502.03591*, 2025.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22035, 2025.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Yuanhong Chen, Fengbei Liu, Hu Wang, Chong Wang, Yuyuan Liu, Yu Tian, and Gustavo Carneiro. Bomd: bag of multi-label descriptors for noisy chest x-ray classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21284–21295, 2023.

Townim F Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W Verjans, and Zhibin Liao. Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2024.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57, 2024.

Jianxiang Dong. Focal loss improves the model performance on multi-label image classifications with imbalanced data. In *Proceedings of the 2nd International Conference on Industrial Control Network And System Engineering Research*, pages 18–21, 2020.

Maria Efimovich, Jayden Lim, Vedant Mehta, and Ethan Poon. Multilabel classification for lung disease detection: Integrating deep learning and natural language processing. *arXiv preprint arXiv:2412.11452*, 2024.

Md Faiyazuddin, Syed Jalal Q Rahman, Gaurav Anand, Reyaz Kausar Siddiqui, Rachana Mehta, Mahalaqua Nazli Khatib, Shilpa Gaidhane, Quazi Syed Zahiruddin, Arif Hussain, and Ranjit Sah. The impact of artificial intelligence on healthcare: a comprehensive review of advancements in diagnostics, treatment, and operational efficiency. *Health Science Reports*, 8(1):e70312, 2025.

Chaowei Fang, Dingwen Zhang, Wen Zheng, Xue Li, Le Yang, Lechao Cheng, and Junwei Han. Revisiting long-tailed image classification: Survey and benchmarks with new evaluation metrics. *arXiv preprint arXiv:2302.01507*, 2023.

Nagarajan Ganapathy, Podakanti Satyajith Chary, Teja Venkata Ramana Kumar Pithani, Pavan Kavati, et al. A multimodal approach for endoscopic vce image classification using biomedclip-pubmedbert. *arXiv preprint arXiv:2410.19944*, 2024.

Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings 23*, pages 757–765. Springer, 2019.

Avani Gupta and PJ Narayanan. A survey on concept-based approaches for model improvement. *arXiv preprint arXiv:2403.14566*, 2024.

Muhammad Shehzad Hanif, Muhammad Bilal, Abdullah H Alsaggaf, and Ubaid M Al-Saggaf. Enhancing multi-label chest x-ray classification using an improved ranking loss. *Bioengineering*, 12(6):593, 2025.

Gregory Holste, Ziyu Jiang, Ajay Jaiswal, Maria Hanna, Shlomo Minkowitz, Alan C Legasto, Joanna G Escalon, Sharon Steinberger, Mark Bittman, Thomas C Shen, et al. How does pruning impact long-tailed multi-label medical image classifiers? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 663–673. Springer, 2023.

Gregory Holste, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, Dongkyun Kim, Trong-Hieu Nguyen-Mau, Minh-Triet Tran, et al. Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. *Medical Image Analysis*, 97:103224, 2024.

Lijie Hu, Tianhao Huang, Huanyi Xie, Xilin Gong, Chenyang Ren, Zhengyu Hu, Lu Yu, Ping Ma, and Di Wang. Semi-supervised concept bottleneck models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2110–2119, 2025.

Haoxu Huang, Samyak Rawlekar, Sumit Chopra, and Cem M Deniz. Radiology reports improve visual representations learned from radiographs. In *Medical Imaging with Deep Learning*, pages 1385–1405. PMLR, 2024.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. *arXiv preprint arXiv:2109.04712*, 2021.

Ta Duc Huy, Sen Kim Tran, Phan Nguyen, Nguyen Hoang Tran, Tran Bao Sam, Anton van den Hengel, Zhibin Liao, Johan W Verjans, Minh-Son To, and Vu Minh Hieu Phan. Interactive medical image analysis with concept-based similarity reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30797–30806, 2025.

Muhammad Irtaza, Arshad Ali, Maryam Gulzar, and Aamir Wali. Multi-label classification of lung diseases using deep learning. *IEEE Access*, 2024.

Lie Ju, Zhen Yu, Lin Wang, Xin Zhao, Xin Wang, Paul Bonnington, and Zongyuan Ge. Hierarchical knowledge guided learning for real-world retinal disease recognition. *IEEE Transactions on Medical Imaging*, 2023.

Lie Ju, Siyuan Yan, Yukun Zhou, Yang Nan, Xiaodan Xing, Peibo Duan, and Zongyuan Ge. Monica: Benchmarking on long-tailed medical image classification. *arXiv preprint arXiv:2410.02010*, 2024.

Nektarios Kalampalikis, Kavya Gupta, Georgi Vitanov, and Isabel Valera. Towards reasonable concept bottleneck models. *arXiv preprint arXiv:2506.05014*, 2025.

Sadman Sadik Khan, Afraz Ul Haque Rupak, Washik Wali Faieaz, Sayma Jannat, Nuzhat Noor Islam Prova, and Amit Kumar Gupta. Advances in medical imaging: Deep learning strategies for pneumonia identification in chest x-rays. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2024.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4230–4238, 2025.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023a.

Injae Kim, Jongha Kim, Joonmyung Choi, and Hyunwoo J Kim. Concept bottleneck with visual concept filtering for explainable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 225–233. Springer, 2023b.

Takumi Kobayashi. Two-way multi-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Jakub Kufel, Michał Bielówka, Marcin Rojek, Adam Mitrega, Piotr Lewandowski, Maciej Cebula, Dariusz Krawczyk, Marta Bielówka, Dominika Kondoł, Katarzyna Bargieł-Łaczek, et al. Multi-label classification of chest x-ray abnormalities using transfer learning techniques. *Journal of Personalized Medicine*, 13(10):1426, 2023.

Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *International conference image analysis and recognition*, pages 546–552. Springer, 2018.

Frank Li, Hari Trivedi, Bardia Khosravi, Theo Dapamede, Mohammadreza Chavoshi, Abdulhameed Dere, Rohan Satya Isaac, Aawez Mansuri, Janice Newsome, Saptarshi Purkayastha, et al. Evaluating vision language models (vlms) for radiology: A comprehensive analysis. *arXiv preprint arXiv:2504.16047*, 2025.

Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3617–3625, 2017.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Fernando Perez-Garcia, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Aravind Sasidharan Pillai. Multi-label chest x-ray classification via deep learning. *arXiv preprint arXiv:2211.14929*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

P Rajpurkar. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv abs/1711*, 5225, 2017.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021.

T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

Divyansh Srivastava, Ge Yan, and Lily Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37: 79057–79094, 2024.

David Steinmann, Wolfgang Stammer, Antonia Wüst, and Kristian Kersting. Object centric concept bottlenecks. *arXiv preprint arXiv:2505.24492*, 2025.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. *arXiv preprint arXiv:2412.07992*, 2024.

Susu Sun, Leslie Tessier, Frédérique Meeuwsen, Clément Grisi, Dominique van Midden, Geert Litjens, and Christian F Baumgartner. Label-free concept based multiple instance learning for gigapixel histopathology. *arXiv preprint arXiv:2501.02922*, 2025.

Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision*, pages 123–138. Springer, 2024.

Nguyen T Thach, Patrick Habecker, Anika R Eisenbraun, W Alex Mason, Kimberly A Tyler, Bilal Khan, and Hau Chan. Muhboost: Multi-label boosting for practical longitudinal human behavior modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 10962–10971, 2023.

Hongmei Wang, Junlin Hou, and Hao Chen. Concept complement bottleneck model for interpretable medical image diagnosis. *arXiv preprint arXiv:2410.15446*, 2024.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 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 16*, pages 162–178. Springer, 2020.

Haotian Xu, Tsui-Wei Weng, Lam M Nguyen, and Tengfei Ma. Graph concept bottleneck models. *arXiv preprint arXiv:2508.14255*, 2025.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023a.

An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023b.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

L Yao, J Prosky, E Poblenz, B Covington, and K Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. arxiv 2018. *arXiv preprint arXiv:1803.07703*, 2018.

Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202, 2018.

Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16786–16793, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023.

Appendix A. Overview of the Selected VLMs

Appendix B. Summary of CBMs Methods

Figure 8 shows the contributions of all concepts across all labels, indicating how they contribute to the final prediction.

Table 3: Summary of selected backbones and their encoders. backbones marked with * were trained or fine-tuned on chest X-rays.

Backbones	**Year**	**Multi-modal**	**Image Enc.**	**Text Enc.**
CLIP	2021	Yes	ViT-B/L	ViT-B/L
BioViL*	2022	Yes	ViT-B	CXR-BERT
BiomedCLIP*	2023	Yes	ViT-B	PubMedBERT
BioViL-T*	2023	Yes	ResNet-50	CXR-BERT
RAD-DINO*	2024	No	ViT-B	—

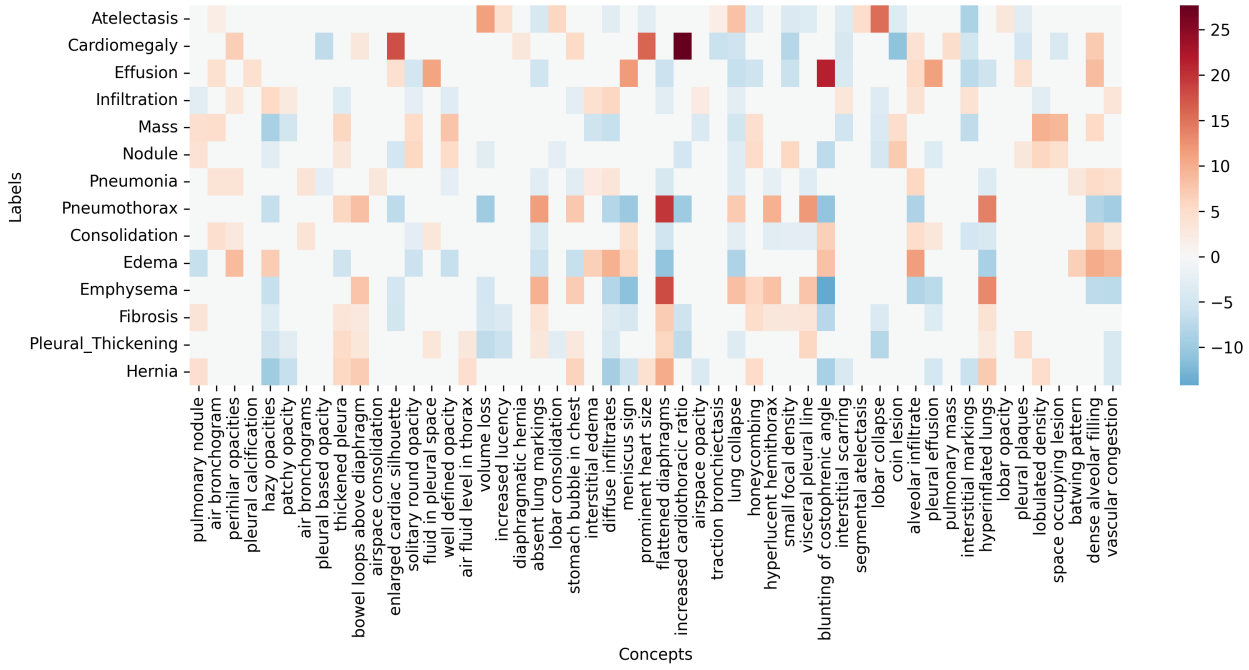

Figure 8: concept-Label contribution Heatmap matrix displaying the contribution of each label.

Table 4: Summary of CBMs, including concept type, dataset, task, and highlights.

Type	Method	Task Type	Dataset(s)	Highlights
Concept-Supervised	CBMs (Koh et al., 2020)	Multi-class	CUB, OAI	Original CBM; allows intervention at concept level.
	ProbCBM (Kim et al., 2023a)	Multi-class	CUB, CelebA	Probabilistic concept embeddings to capture uncertainty.
	CREAM (Kalampalikis et al., 2025)	Multi-class	CUB, CelebA, FMNIST	Models concept-concept and concept-task relationships; uses side-channel.
	CSR (Huy et al., 2025)	Multi-class	VinDr-CXR, TBX11, ISIC	Patch-level prototypes; spatially localized explanations for medical imaging.
	GraphCBM (Xu et al., 2025)	Multi-class	CUB, Flower102, HAM10000, Cifar-10, Cifar-100, CheXpert	More concept structure information for interpretability and use of latent concept graphs for more effective interventions.
	SSCBM (Hu et al., 2025)	Multi-class	CUB, AwA2, WB-Catt	Joint training on labeled and unlabeled data; alignment loss for concepts.
Label-Free	LF-CBM (Oikarinen et al., 2023)	Multi-class	CIFAR, CUB, ImageNet, Places365	Transforms standard networks into CBMs; scalable construction.
	LaBo (Yang et al., 2023)	Multi-class	CIFAR, CUB, ImageNet, HAM10000	Automated concept generation; high performance without manual annotation.
	Visual-Filtered CBM (Kim et al., 2023b)	Multi-class	HAM10000, ImageNet, COCO	Filters LLM concepts by visual relevance; improves performance
	BotCL (Wang et al., 2023)	Multi-class	MNIST, CUB, ImageNet	Learns human-understandable concepts without supervision.
	OpenCBM (Tan et al., 2024)	Multi-class	CUB, BDD-OIA	Open vocabulary concepts; allows adding/removing concepts post-training.
	Object-Centric CBM (Steinmann et al., 2025)	Multi-class	COCOLogic, PASCAL-VOC, SUN397	Object-centric representation; supports multi-label tasks.
	Concept MIL (Sun et al., 2024)	Multi-class	Camelyon16, PANDA	Whole-slide histopathology; interpretable predictions without manual annotation.
	CBM-RAG (Alam et al., 2025)	Multi-class	NIH CXR dataset	Multi-agentic radiology report generation; interpretable classification.
	Concept Complement CBM (Wang et al., 2024)	Multi-class	Derm7pt, Skincon, Breast US, LIDC-IDRI	Learns new concepts complementing existing ones; medical image interpretation.
	LLM-CBM (Yan et al., 2023a)	Multi-class	CUB, CIFAR-10, CIFAR-100, Food-101, Flower, Oxford-Pets, Stanford-Cars, ImageNet	Leverages a concise subset of descriptive visual concepts from LLM-generated concepts
	AdaCBM (Chowdhury et al., 2024)	Multi-class	HAM10000, BCCD, DR	Introduces an adaptive module between CLIP and the CBM to improve feature alignment, boosting classification performance while preserving interpretability