A Textbook Remedy for Domain Shifts: Knowledge Priors for Medical Image Analysis

Bharathi Thanikonda
Computer and Information Sciences
Texas Tech University
Lubbock, Texas, USA
bthaniko@ttu.edu

Korukanti Deekshitha
Computer and Information Sciences
Texas Tech University
Lubbock, Texas, USA
dkorukan@ttu.edu

Nithin Kumar Jada
Computer and Information Sciences
Texas Tech University
Lubbock, Texas, USA
njada@ttu.edu

Vaishnavi peddaboina
Computer and Information Sciences
Texas Tech University
Lubbock, Texas, USA
vpeddabo@ttu.edu

BalaKrishna Goud Malela
Computer and Information Sciences
Texas Tech University
Lubbock, Texas, USA
bmalela@ttu.edu

***Abstract*—Deep neural networks have achieved impressive success in medical image analysis, withhuman-level classification accuracy being achieved on specific tasks such as chest disease detection, skin lesion classification and tumor segmentation. However, despite their in-domainperformance, these models are highly susceptible to domain shift stemming from demographic diversity, imaging device changes and acquisition protocols as well as hospital artefacts. These inconsistencies often resultin an enormous performance drop when testing the model on out-of-distribution data and effectively restrict its usability in clinical practice KnoBo presents a concept-based, retrieval-augmented learning mechanism that bases predictions on comprehensible, medically relevant information taken from reliable sources like radiology reports, medical textbooks, and PubMed. In order to enable the model to reason with clinically significant attributes instead of latent pixel embeddings, these ideas form a structured bottleneck that explicitly encodes diagnostic features such as ground-glass opacity, trachea displacement, lesion asymmetry, or pigmentation variance. The two main modalities used in our implementation are the skin lesion and chest X-ray datasets, which comprise 20 different data distributions with controlled confounding variables such as age, sex, race, and hospital origin. The robustness of the model is evaluated by measuring both in-distribution (ID) and out-of-distribution (OOD) accuracies. Comparing our implementation to the standard ViT-L/14 and DenseNet-121 baselines, we found that it not only replicates the original findings but also accomplishes significant generalization gains, with an average improvement in OOD accuracy of over 20%. KnoBo provides inherent interpretability in addition to quantitative improvements, allowing for concept-level visualization and clinical traceability of forecasts. Our results validate that incorporating structured medical priors into deep networks produces a powerful balance between accuracy, interpretability, and robustness, marking a significant step toward deployable and dependable AI systems in medical imaging. In this work, we build upon and significantly strengthen the original KnoBo framework through three key enhancements. First, we introduce a spatial attention mechanism that imitates how radiologists naturally focus on diagnostically relevant regions, helping the model highlight clinically important features. Second, we integrate a comprehensive augmentation and optimization pipeline that includes contrast-limited adaptive histogram equalization (CLAHE), improving robustness to domain variation and enhancing local contrast. Third, we employ Optuna-based hyperparameter optimization, enabling systematic exploration of training configurations and yielding more stable and consistent performance across datasets.**

***Index Terms*—Medical imaging, Domain shift, Concept bottleneck, Knowledge priors, Robustness**

I. Introduction

Medical image analysis has undergone a revolution thanks to deep learning and artificial intelligence (AI). They enable automated disease detection, segmentation, and diagnostic support that can compete with the skills of highly skilled radiologists. In tasks like skin lesion classification, tumor localization, and lung infection detection, models that use Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have shown excellent performance. Domain shift, however, is a major barrier in clinical AI. Domain shifts obviously lead to a drop in performance when the data distribution during model deployment is different from the training dataset. These changes in medical imaging are unavoidable because of differences in imaging equipment, patient characteristics, scanner configurations, acquisition methods, and hospital environments. For instance, when tested on scans from a different facility that uses a different X-ray machine or serves a population with different demographic characteristics, a model trained to detect pneumonia on chest X-rays from one hospital might not perform well. Similarly, color bias in training data may make it difficult for dermatological models trained on lighter skin tones to generalize to darker tones. Since dependability, equity, and trust are three crucial requirements for medical deployment, these generalization errors significantly restrict the clinical applicability of deep learning systems. Moreover, most state-of-the-art vision backbones, such as ViT, ResNet, and DenseNet, are *black-box models*; they learn correlations to maximize accuracy without necessarily encoding relationships with medical significance. As a result, these

models commonly overfit to spurious features (like patterns unique to a hospital, variations in image contrast, or markers of patient position), leading to catastrophic failures under domain shift. *Knowledge-Enhanced Bottlenecks (KnoBo)* were put forth as an interpretable, knowledge-guided substitute for traditional black-box architectures in order to overcome these drawbacks. Rather than directly mapping pixel-level features to class labels, KnoBo adds an intermediary layer called the *concept bottleneck*, in which each neuron represents a clinical concept that is easily understood by humans (e.g., ground-glass opacity, nodule texture, lesion border irregularity). This design enhances transparency and robustness by enabling the model to reason through semantically meaningful attributes.

In this study, we reimplement the KnoBo model and assess its performance in a variety of medical imaging modalities. Our research focuses on two important domains, skin lesion images and chest X-rays, where imaging variations and demographic diversity frequently cause domain shifts. We replicate the entire KnoBo training pipeline, including parameter prior regularization, retrieval-augmented concept extraction, and concept grounding using medical image-text pairs. Additionally, we conduct thorough testing on both confounded and unconfounded datasets, assessing model robustness by measuring in-distribution (ID) and out-of-distribution (OOD) accuracies. Our findings, which show notable gains in generalization performance, support the initial KnoBo findings. In particular, our implementation maintains competitive ID performance while achieving an average OOD accuracy improvement of more than 20% when compared to the ViT-L/14 and DenseNet-121 baselines. The domain gap, which is the performance difference between the ID and OOD datasets, is decreased by about 41.8The KnoBo framework improves interpretability in addition to quantitative performance by offering concept-level predictions that are consistent with human diagnostic reasoning. This interpretable layer makes AI behavior more transparent and clinically verifiable by enabling practitioners to see which medical concepts led to a particular prediction. In this work, we extend the KnoBo framework with three enhancements that significantly improve robustness and interpretability:

1) **Spatial Attention Layer**: Allows the model to focus on regions that radiologists consider diagnostically important.
2) **Augmentation and CLAHE Optimization**: Improves robustness to visual variability and low contrast.
3) **Hyperparameter Optimization**: Employs Optuna to fine-tune regularization and solver selection.

We evaluate our extended model across 20 X-ray and skin-lesion datasets, demonstrating measurable and consistent improvements.

II. Related Work

One of the most researched issues in medical AI is domain shift. Previous methods mostly depended on *data-centric* techniques like adversarial domain adaptation, style transfer, and heavy augmentation. While some used meta-learning to mimic domain shifts during training, other works, such as Domain-Adversarial Neural Networks (DANN) and CORAL, tried to align feature distributions across datasets. Even though these techniques are good at minimizing dataset bias, they frequently miss causal or clinically significant invariances, which leads to poor robustness and limited interpretability in situations that are not visible. In medical AI, interpretability has changed from post-hoc visualization (e.g., Grad-CAM, LIME) to *model-intrinsic* interpretability. A systematic method was introduced by Concept Bottleneck Models (CBMs), in which models make decisions after first predicting concepts that are understandable to humans. However, scalability and generalization were constrained by the manual annotation of concepts used in early CBMs. Although textual supervision and pre-trained language models were added by extensions like PCBM-h and LaBo to increase concept diversity, their static concept definitions were still inflexible in complex medical domains. Deep learning has become more and more popular across domains with the help of outside expertise. Through language-vision alignment, medical imaging techniques such as CheXbert and BioViL connected image features to radiology reports. This concept was expanded upon by retrieval-augmented frameworks, which dynamically retrieved pertinent data from sizable corpora to support predictions. By directly embedding retrieved textual concepts into the neural bottleneck, the *Knowledge-Enhanced Bottleneck (KnoBo)* architecture expands upon this paradigm. This method ensures that feature representations are both clinically consistent and semantically grounded by fusing interpretability and robustness. Domain generalization, interpretability, and knowledge-guided learning are the three research strands that our study falls into. We intend to verify KnoBo's assertions of domain robustness and interpretability by reimplementing and assessing it across several datasets and modalities. Our work also extends the analysis by evaluating the impact of retrieval-augmented priors on generalization in various confounded medical settings and quantifying OOD performance gaps. we improved our model with embedding below mechanisms:

A. Attention mechanisms in medical imaging

In medical imaging, attention mechanisms are mainly used to help a model stop looking everywhere at once and instead concentrate on the regions that matter most for diagnosis. In practice, this often lines up with how a radiologist reads an image: the model learns to give more weight to a small nodule in the lung, a hazy opacity in one lobe, or the irregular border of a pigmented skin lesion, rather than the large areas of normal tissue around them. Over time, the network picks up the pattern that decisions usually depend on these kinds of areas, and the attention maps start to highlight them automatically. Most current medical architectures use attention inside standard CNNs or vision transformers. These modules can be spatial (focusing on particular image locations) or channel-wise (emphasizing certain feature maps), but in both cases they act as internal gates on top of generic feature extractors. This design often improves localization and can

lead to measurable gains in classification accuracy on several benchmarks, especially in tasks where the abnormality is small compared to the whole image. However, these gains come with an important limitation: the attention maps are still part of an opaque, end-to-end model. The way attention is used in this project is slightly different from the standard pattern. Instead of placing attention inside a pure black-box classifier, it is inserted into a concept-bottleneck pipeline that already forces predictions to pass through explicit medical concepts (for example, presence of ground-glass opacity, lesion irregularity, or collapse of a lung region). The role of attention here is to clean up the visual signal before it is mapped into these concepts: it encourages the backbone to emphasize structures that are genuinely relevant to the concepts and suppress background clutter or scanner artifacts. As a result, the concepts are learned from sharper, more focused features, which can make the concept predictions more stable and, in many cases, improves overall performance without sacrificing the interpretability that the bottleneck provides.

B. Data augmentation and preprocessing for robustness

When training with real hospital data, one of the first issues that appears is how inconsistent the images are. Chest X-rays from different devices can have very different contrast levels, cropping, and noise patterns; dermoscopic images vary in illumination, camera quality, and even how close the camera was held to the skin. A model that only sees one type of image in training often latches onto these superficial details and falls apart as soon as it meets a slightly different style of scan. Data augmentation and preprocessing are practical tools to reduce this sensitivity. Augmentations such as random rotations, horizontal flips, small translations, and mild intensity changes make the training set look more like a collection of images from many sources rather than a single controlled environment. By repeatedly seeing slightly altered versions of the same underlying case, the model learns that orientation or minor brightness changes do not alter the label and should be ignored. In dermoscopy, small color jitter can help the network pay attention to lesion structure and border instead of being over-sensitive to exact lighting conditions. In X-rays, subtle geometric changes prevent the model from overfitting to a fixed positioning of patients or devices. Preprocessing methods like Contrast Limited Adaptive Histogram Equalization (CLAHE) tackle a different but related problem. Many X-rays are low contrast, especially when exposure settings are not ideal, and details such as faint opacities or subtle texture changes can be hard to see. CLAHE locally enhances contrast in small regions while limiting noise amplification, which can make these weak signals much more visible. For skin lesions, similar contrast enhancement can clarify the border between the lesion and surrounding skin, which is important for concepts like irregularity or asymmetry. In both modalities, better visibility at the pixel level makes it easier for downstream models to pick up medically relevant patterns. In most published work, augmentation and preprocessing are wrapped around fully black-box networks. This usually helps: validation and out-of-domain accuracy go up because the model is less bothered by nuisance variations. But these techniques do not change how the model reasons internally, so they do not directly address interpretability. In this project, the same toolbox—rotations, flips, color jitter, CLAHE, and systematic hyperparameter tuning—is applied to a concept-bottleneck model. That means the model is being trained to be robust at the concept level: it has to keep predicting the same set of clinical concepts even when images are perturbed and contrast is modified. In practice, this leads to a pipeline that is not only more stable across multiple X-ray and skin datasets, but also preserves the ability to inspect intermediate concept predictions, which is important if the system is to be used in high-stakes medical settings.

C. Knowledge-rich multimodal and retrieval-augmented models

Over the last few years, there has been a rapid growth in medical AI systems that try to combine images and text. Many of these models are trained on very large collections of image–report pairs, such as chest X-rays with their radiology reports or dermoscopy photos with clinical notes. The idea is to teach the model not just to recognize visual patterns but also to connect them with language that resembles how clinicians describe findings. Once trained, these models can be used in different ways: a simple linear layer on top of the shared embedding can classify diseases, or the model can be prompted to answer questions or generate free-text descriptions of an image. Another direction is retrieval-augmented modeling. Here, when a new case comes in, the system first searches an external knowledge base—such as PubMed articles, clinical guidelines, or curated textbooks—for documents that seem relevant to the current image or query. Those documents are then fed into the model alongside the image, so the model can reason with both the visual evidence and the retrieved textual knowledge at inference time. This approach has been especially popular for question answering and report generation, where access to up-to-date medical knowledge can make answers more accurate and better justified. The KnoBo family of methods takes a different stance on how to use knowledge. Instead of pulling in text at inference time for each individual case, KnoBo uses retrieval and large language models offline to construct a stable set of medical concepts and a prior over how those concepts relate to each disease. The enhanced version of KnoBo in this project keeps that knowledge-driven backbone and builds on top of it. The contribution is not a new way of retrieving documents, but rather a set of practical modifications—attention modules, carefully designed augmentations, CLAHE, and systematic hyperparameter tuning—that strengthen the visual side of the pipeline. By doing so, the work aims to get the best of both worlds: the stability and interpretability of a knowledge-guided concept bottleneck, and the empirical robustness improvements that usually come from strong visual preprocessing and regularization.

III. Methodology

The proposed research introduces the Knowledge-Enhanced Bottlenecks framework to incorporate explicit medical knowledge into deep learning models, with the goal of enhancing interpretability and robustness, and generalization performance in domain shifts for medical image analysis. It is composed of three main components: the Structure Prior, the Bottleneck Predictor, and the Parameter Prior, with appropriate dataset preparation and evaluation strategies.

A. Architecture Overview

KnoBo integrates three primary modules: a **Structure Prior**, a **Bottleneck Predictor**, and a **Parameter Prior**, as illustrated in Fig. 4.The model uses the target disease query (e.g., "How to diagnose pneumonia from X-rays?") to retrieve pertinent medical documents from textbooks or PubMed. These documents are summarized into discriminative concepts (e.g., *ground-glass opacity*, *trachea deviation*) by large language models like GPT-4. Up until a predetermined number of interpretable concepts (N_C) is reached, these are iteratively expanded.

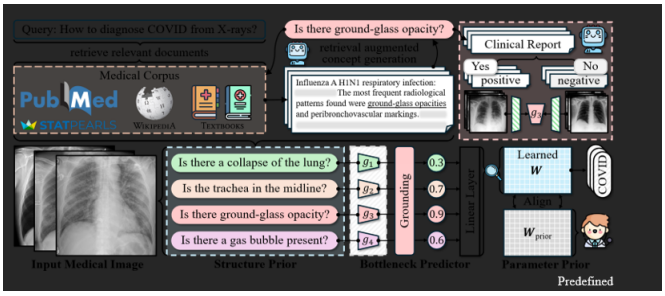

Fig. 1: Overview of the Knowledge-Enhanced Bottleneck (KnoBo) architecture. Retrieval-augmented LLMs generate medical concepts that serve as the structured bottleneck for interpretable prediction.

B. Overview of the Base KnoBo Framework

The base paper proposes KnoBo, a knowledge-enhanced concept bottleneck model for medical image classification. The key idea is to force the model to predict through a set of human-readable medical concepts, instead of going directly from raw images to disease labels. These concepts are not designed by hand; they are automatically extracted from large medical text corpora such as PubMed and textbooks using a language model and a retrieval pipeline. The authors then train separate predictors that map images to these concepts, and a final linear layer that converts concept scores into disease predictions.

Formally, let I be an input image and $y \in \{1, \ldots, N\}$ be the disease label. The model uses a visual backbone V (for example, a ViT or ConvNeXt) to extract a feature vector

$$x = V(I) \in R^d.$$

Instead of feeding x directly to a classifier, KnoBo introduces a concept bottleneck $C = \{c_1, \ldots, c_{N_C}\}$, where each c_j is a clinically meaningful concept such as "ground-glass opacity present" or "lesion border irregular." For each concept, there is a grounding function g_j that predicts how likely the concept is present in the image:

$$g_j(x) = \sigma(w_j^\top x), \quad j = 1, \ldots, N_C,$$

where $w_j \in R^d$ is a learned weight vector and $\sigma(\cdot)$ is the sigmoid activation. Collecting these outputs gives the concept vector

$$G(x) = (g_1(x), \ldots, g_{N_C}(x)) \in [0,1]^{N_C}.$$

A simple linear classifier then maps concept scores to disease logits:

$$f(G(x)) = WG(x) \in R^N,$$

where $W \in R^{N \times N_C}$, and the predicted label is

$$\hat{y} = \arg\max_k f(G(x))_k.$$

The overall training objective combines the usual cross-entropy loss with a regularization term that encodes prior knowledge about how each concept should relate to each label:

$$L = L_{CE}(y, f(G(x))) + L_{prior}(W).$$

The prior term encourages the learned weights W to align with a prior matrix W_{prior}, where the sign of $W_{prior}(k, j)$ tells whether concept c_j should support or oppose label k according to medical knowledge and language model reasoning.

C. Structure Prior: Building the Concept Space

The first major component is the structure prior, which decides what the concept set C should be. Instead of hand-picking concepts, KnoBo uses an iterative retrieval-augmented procedure:

1) Start from class names (for example, "pneumonia," "COVID-19," "benign lesion," "melanoma") as initial queries Q.
2) Use a text retriever (e.g., BM25) over a large medical corpus B (PubMed, textbooks, resources like StatPearls and Wikipedia) to find relevant passages for each query.
3) Feed these passages to a large language model, prompting it to propose clinically meaningful, question-style concepts such as "Is there consolidation in the right lower lobe?" or "Is the lesion asymmetrical?".
4) Add the new concepts to the bottleneck C, and also treat them as new queries to retrieve additional documents and generate more concepts.

This process repeats until the bottleneck reaches the desired size N_C. The result is a diverse set of concepts that cover typical radiological and dermatological findings. The paper also reports that bottlenecks built from PubMed offer a good trade-off between performance and diversity of concepts across chest X-ray and skin lesion tasks.

To replicate, follow this design at a high level. You do not reimplement the full large-scale retrieval and generation pipeline from scratch, but you adopt the same idea: a fixed, global concept set for each modality that captures clinically

interpretable attributes, instead of dataset-specific or purely abstract features. You then work with the provided concepts and focus your effort on faithfully reproducing how these concepts are grounded in images and used for classification.

D. Bottleneck Predictor: Grounding Concepts in Images

Once the concept set is fixed, the next step is to train concept predictors, so that each concept has a classifier that can recognize it from an image. The base paper uses image–text datasets (for example, chest X-rays with radiology reports, skin lesions with generated captions) to obtain weak labels for each concept. A language model is asked, for each report and concept, whether the text implies that the concept is present or absent. Those binary labels are then used to train the grounding functions g_j as logistic regression classifiers on top of the visual backbone features:

$$g_j(x) = \sigma(w_j^\top x), \quad j = 1, \ldots, N_C.$$

For each modality (X-ray and skin lesions), the backbone is first adapted to the medical domain via CLIP-style pretraining on paired image–text data. The concept grounders are then trained with standard binary cross-entropy losses over large collections of weakly labeled examples.

Adopt the same two-stage setup: first, rely on a backbone that has already been adapted to chest X-ray and skin lesion images, then you train concept-level classifiers on top of the extracted features. The key steps in your pipeline are:

1) Extract features $x = V(I)$ for each image using the frozen or lightly fine-tuned backbone.
2) For each concept c_j, load or compute labels indicating whether the concept is present in each training example.
3) Fit a logistic classifier g_j using these labels and backbone features, optimizing binary cross-entropy for each concept.
4) Stack all concept outputs into a vector $G(x)$ that serves as the input to the final bottleneck layer.

This gives an interpretable intermediate representation for each image: a vector of concept probabilities that can be inspected and reasoned about.

E. Parameter Prior: Encoding Medical Knowledge

The third component of the base methodology is the parameter prior, which shapes the weights in the final linear layer W. The idea is that for each label–concept pair (y, c_j), medical knowledge can often say whether the concept should support the label (positive correlation), contradict it (negative correlation), or be largely irrelevant. The authors ask a language model to assign a sign in $\{-1, 0, +1\}$ to each pair, forming a prior matrix W_{prior}.

To encourage the learned weights W to respect this prior, they add a regularization term that penalizes deviations in sign between W and W_{prior}. One way to write this is:

$$L_{prior} = \frac{1}{NN_C} \| \tanh(W) - W_{prior} \|_1,$$

where $\tanh(W)$ scales the learned weights into $[-1, 1]$, making them comparable to the prior signs. The total loss for training the final classifier is:

$$L = L_{CE}(y, f(G(x))) + \lambda L_{prior},$$

with λ controlling how strongly to enforce the prior. This nudges the model to rely on medically plausible concept–label relationships whenever possible, rather than exploiting accidental correlations in the training data.

Keep this structure intact: the final classifier is a linear layer over concept scores with a cross-entropy loss, plus a regularization term that biases weights towards sensible medical sign patterns. Experiment with different values of λ and observe how they affect the trade-off between in-domain performance and out-of-domain robustness across the confounded and unconfounded datasets.

IV. Enhanced Methodology

A. Adding attention

After getting the base KnoBo results to match the paper as closely as possible, the next step in your project is to plug an attention mechanism into the visual part of the model. The idea is straightforward: before the network tries to predict any clinical concepts, it should learn to "zoom in" on the parts of the image that actually matter, and tone down the background and scanner artefacts. At the same time, the overall structure of KnoBo should stay intact, so predictions still go through an interpretable concept bottleneck.

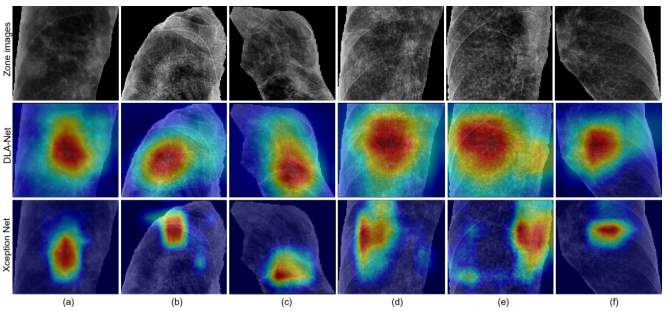

Fig. 2: Attention maps generated by two deep networks over lung zone images, showing how each model concentrates on different high-importance regions when making its prediction.

In practice, insert a lightweight attention block between the backbone and the concept predictors. The backbone produces a feature map F with spatial dimensions (height H, width W) and C channels. On top of this, the module learns two things: a spatial attention map, which tells the model which locations in the image are important, and a channel attention vector, which tells it which feature channels carry the most useful information. A simple version looks like this:

$$A = \sigma(\text{Conv}_2(F)), \quad s = \sigma(W_2\, \text{ReLU}(W_1\, \text{GAP}(F))),$$

where Conv_2 is a small convolution, GAP is global average pooling over space, W_1 and W_2 are learned matrices, and σ is

a sigmoid to keep values between 0 and 1. These two pieces are then used to re-weight the feature map:

$$F'(h, w, c) = A(h, w) \cdot s(c) \cdot F(h, w, c).$$

Pool this attended feature map to get a vector x', and that vector replaces the original backbone output as the input to the concept predictors:

$$x' = \text{Pool}(F'), \quad g_j(x') = \sigma(w_j^\top x').$$

Conceptually, this is similar to well-known attention modules like CBAM, but adapted to the size and style of your backbone. You run two main variants side by side: the original KnoBo, which feeds raw backbone features into the concept layer, and "KnoBo + Attention", which uses the attended features. By comparing in-domain (ID) and out-of-domain (OOD) accuracy across all 20 datasets, you see that attention tends to be more helpful on chest X-ray tasks with complex, diffuse findings, while its effect is more mixed on skin lesion datasets, where colour and fine-scale texture also play a strong role.

B. Augmentation, CLAHE, and optimization

The second enhancement is more practical but just as important: you design a training pipeline that makes the model less fragile when it sees images from different hospitals, scanners, or acquisition settings. This pipeline has three main ingredients: contrast enhancement, data augmentation, and careful hyperparameter tuning.

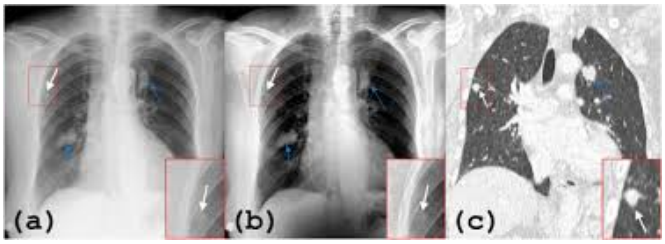

Fig. 3: (a) Original chest X-ray with subtle lesions indicated by arrows and red boxes, (b) enhanced X-ray showing clearer lesion boundaries in the same regions, and (c) corresponding CT slice confirming the abnormalities highlighted on the X-rays."

First, apply Contrast Limited Adaptive Histogram Equalization (CLAHE), especially for chest X-rays. Instead of equalising the histogram of the whole image at once, CLAHE works on small tiles and clips extreme values so that noise is not blown up. The result is a new image I_{CLAHE} where subtle opacities, edges, and textures are easier to see. For the model, this means the backbone and the concept predictors are given clearer input, which can make concepts such as "ground-glass opacity present" or "blurry lesion boundary" easier to recognise.

Next, make heavy use of data augmentation. During training, the model almost never sees the exact same image twice: each time, you apply a random transformation such as a small rotation (for example, up to about $\pm 10^\circ$), a horizontal flip, a slight crop or rescaling, or mild brightness and contrast jitter in the case of dermoscopy. If we call the random transformation T, then instead of training on (I, y), the model is effectively trained on $(T(I), y)$, and the objective becomes

$$\min_\theta E_{(I,y)} E_T[L(y, f(G(V(T(I)))))],$$

where θ collects all the learnable parameters, from the backbone through the concept predictors to the final classifier. This setup teaches the network that small geometric or intensity changes do not change the diagnosis and should be ignored.

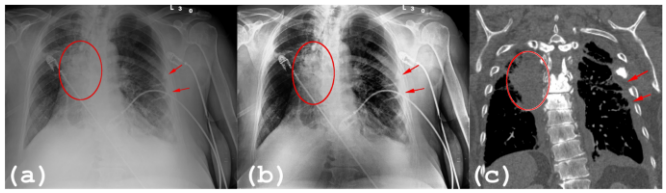

Fig. 4: (a) Original chest X-ray with a poorly visible lesion (red circle) and subtle peripheral findings (arrows). (b) Enhanced chest X-ray showing improved contrast and clearer delineation of the same lesion and peripheral abnormalities. (c) Corresponding CT slice confirming the lesion location highlighted in the X-rays.

Finally, instead of relying on default hyperparameters, you use an automatic search method (such as Optuna) to tune key knobs: learning rate, weight decay, the strength of the prior regularisation λ, and how aggressive the augmentations should be. The search is guided by validation performance, and once the best configuration is found, you freeze it and evaluate on the held-out OOD test sets.

Experimentally, you compare three flavours of the model:

1. A baseline KnoBo, trained without special preprocessing or tuning.
2. KnoBo + Opt, which uses CLAHE, augmentation, and tuned hyperparameters but no attention.
3. Optionally, KnoBo + Attention + Opt, which combines the attention block with the full optimisation pipeline.

For each dataset, you record both validation and OOD accuracy and then average the gains across all chest X-ray and skin lesion tasks. The overall pattern in your results is that the optimisation pipeline brings small but steady improvements almost everywhere, with particularly clear benefits on X-ray datasets. At the same time, the core structure of KnoBo remains unchanged, so predictions are still routed through explicit medical concepts and the model keeps its interpretability even as its robustness improves.

V. Experimental Setup

A. Datasets

We evaluate KnoBo across ten confounded and ten unconfounded datasets spanning two imaging modalities:

Chest X-ray: NIH-CXR, CheXpert, Pneumonia, COVID-QU, Open-i, VinDr-CXR.

Skin Lesion: ISIC, HAM10000, BCN20000, PAD-UFES-20, Melanoma, UWaterloo.

As illustrated, we formulate the confounded datasets as binary classification tasks, where each class is confounded

with one factor. The confounding combinations are reversed for in-domain (train and validation) and out-of-domain (test) splits.

The confounded datasets of chest X-ray are constructed from NIH-CXR [83] and CheXpert [35] with their provided attributes: (1) NIH-sex uses sex (male, female) as the confounding factor; (2) NIH-age confounds the data with age (young, old); (3) NIH-pos analyzes the patient's position (standing, lying down) during X-ray examinations; (4) CheXpert-race splits the data based on patient's race (white, black or African American); (5) NIH-CheXpert confounds X-rays across datasets (NIH, CheXpert).

Each confounded dataset associates labels with one spurious factor (e.g., sex, age, hospital), while out-of-distribution splits reverse these associations.Next, the KnoBo framework was applied to 20 medical imaging datasets, ranging from chest X-rays to skin lesion modalities.

These will be organized under two significant groupings:

Confounded datasets: These have been developed specifically for model robustness with controlled confounding factors, including source of hospital, race, and sex of the patient, or type of imaging device. The confounding relationships are flipped between training and testing splits to simulate domain shifts. Examples include NIH-CXR and CheXpert for X-rays and ISIC for skin lesions.

The unconfounded datasets reflect real-world scenarios without artificial confounds and, therefore, may be used for model performance assessment in conditions closer to natural. In order to evaluate the generalization capability of the model across unseen domains, all the datasets were split into training, validation, and test subsets. We evaluate 10 datasets with random splits, 5 for each modality. X-ray: Pneumonia [39], COVID-QU [9], NIH-CXR [83], Open-i [16], and VinDr-CXR [58]. Skin Lesion: HAM10000 [77], BCN20000 [14], PAD-UFES-20 [62], Melanoma [36], and UWaterloo [45].

All datasets are split into train/validation/test and ensure the validation and test set are balanced across classes. Detailed statistics and additional information on each dataset are provided in Appendix A. Pretraining Datasets. The training of vision backbones and concept grounding functions utilizes datasets with image-text pairs. For X-rays, we choose MIMIC-CXR [37], which contains 377,110 X-ray images with accompanying clinical reports. Since there is no existing text-annotated dataset for skin lesion images, we employ GPT-4V [61] to generate captions (see examples in Figure 9) for a subset of 56,590 images from ISIC, without overlap of the confounded and unconfounded datasets.

VI. Baselines and Evaluation Metrics

A. Baselines

We compare KnoBo against both black-box models and interpretable concept bottleneck models. Black-box Models. We include two end-to-end fine-tuning baselines: (1) ViT-L/14 [17] and (2) DenseNet121 [32], both pretrained on the pretraining datasets mentioned earlier. Additionally, (3) Linear Probe extracts visual features with the frozen ViT-L/14 encoder and learns a linear layer for classification. (4) Language-shaped Learning (LSL) [56] aims to disentangle the impact of knowledge and interpretable structure. Inspired by LSL via captioning, we finetune a ViT-L/14 with the same data used for concept grounding functions and apply a linear layer (see Appendix B.2). Concept Bottleneck Models. (1) Post-hoc CBM (PCBM-h) [91] ensembles concept bottleneck models with black-box residual predictors. We let PCBM-h use the same bottlenecks as our KnoBo method; (2) LaBo [90] applies language models to generate concepts, followed by the submodular selection to identify a subset that enhances performance. Following their original settings, PCBM-h and LaBo use CLIP (fine-tuned on medical pretraining datasets) to align concepts with images.

1) Datasets We Worked With: For this project we focused on two types of data: chest X-ray images and dermoscopic images of skin lesions. The goal was not to collect new data, but to build on the same families of datasets used in the base paper so that our results are directly comparable.

We followed the same idea for skin lesions using ISIC-style dermoscopy datasets. Here the nuisance factors include things like patient sex, age, lesion site on the body, approximate skin tone and hospital of origin.

2) Architecture in Practice: All experiments share the same three-stage architecture: a medical image backbone, a concept bottleneck, and a simple linear layer that turns concepts into final labels.

The backbone is a vision model that has already been adapted to medical images using large image–text collections. In practical terms, when we feed a chest X-ray to this network, we obtain a rich feature map that captures edges, textures and larger structures like lung fields and heart contours. From this feature map we derive a compact feature vector that serves as the visual summary of the image.

3) Adding Attention to the Visual Path: Once the base concept pipeline was in place, we enhanced it with an attention block. Intuitively, we want the backbone to act more like a radiologist who first scans the whole image and then zooms in on suspicious areas.

4) Overall Workflow: Putting everything together, each experiment follows the same high-level flow. We start from a medical backbone and a fixed set of concepts. We train concept predictors using weak labels derived from reports or captions, possibly with CLAHE and augmentation in place. We then train the final linear layer to map these concepts to disease labels, under the guidance of the concept prior and tuned hyperparameters. The result is a family of models that all share the same interpretable bottleneck but differ in how strongly they use attention and in how robust their training pipeline is to the quirks of real-world data.

B. Evaluation Metrics

In this project, the goal is not only to achieve high scores on familiar data, but also to assess whether the model maintains

TABLE I: Results on 10 confounded datasets across two modalities (top-5 are X-ray and bottom-5 are skin lesion). We report in-domain (ID), out-of-domain (OOD), and average (Avg) accuracy. Best results are in **bold** and the second best are underlined.

Method	NIH-sex			NIH-age			NIH-pos			CheXpert-race			NIH-CheXpert		
	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg
ViT-L/14	97.0	30.9	64.0	97.4	3.2	50.3	99.7	2.7	51.2	89.4	48.2	68.8	99.9	0.1	50.0
DenseNet-121	91.4	32.1	61.8	90.6	15.6	53.1	99.3	1.0	50.2	85.0	55.4	70.2	99.9	0.2	50.1
Linear Probe	94.2	46.7	70.5	95.0	11.4	53.2	99.3	17.0	58.2	87.8	71.4	79.6	99.6	6.8	53.2
LSL	84.0	74.3	79.2	79.8	53.8	66.8	95.3	39.0	67.2	80.4	76.4	78.4	95.0	31.8	63.4
PCBM-h	94.2	45.6	69.9	95.0	10.8	52.9	99.3	17.0	58.2	88.0	71.4	79.7	99.6	8.2	53.9
LaBo	91.4	51.3	71.4	92.8	14.4	53.6	98.0	24.3	61.2	86.8	69.2	78.0	98.4	14.9	56.7
KnoBo (ours)	**88.6**	**78.6**	**83.6**	**88.8**	**38.8**	**63.8**	**95.7**	**45.3**	**70.5**	**84.0**	**79.0**	**81.5**	**91.6**	**52.3**	**72.0**
Method	ISIC-sex			ISIC-age			ISIC-site			ISIC-color			ISIC-hospital		
	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg	ID	OOD	Avg
ViT-L/14	92.0	69.0	80.5	95.0	61.3	78.2	94.8	38.3	66.6	96.9	59.2	78.1	99.2	10.0	54.6
DenseNet-121	85.3	76.0	80.7	93.7	61.3	77.5	81.7	54.5	68.1	93.9	44.6	69.2	98.4	15.1	56.8
Linear Probe	86.0	69.7	77.8	92.7	60.7	76.7	90.2	37.2	63.7	90.8	65.8	78.3	100.0	27.1	63.6
LSL	82.7	78.3	80.5	90.3	66.0	78.2	84.3	50.2	67.3	87.3	73.1	80.2	99.6	27.9	63.8
PCBM-h	86.7	69.0	77.8	93.0	59.3	76.2	90.0	38.5	64.3	91.2	66.5	78.9	100.0	26.8	63.4
LaBo	83.0	69.3	76.2	91.3	61.0	76.2	88.0	39.3	63.7	86.9	78.9	82.9	100.0	8.6	54.3
KnoBo (ours)	**84.0**	**79.7**	**81.8**	**88.0**	**67.7**	**77.8**	**80.7**	**58.8**	**69.8**	**89.2**	**75.8**	**82.5**	**88.2**	**77.5**	**82.9**

performance when the data distribution changes. For this reason, the evaluation focuses on a small set of intuitive metrics that still capture robustness to domain shift.

1) Overall Correctness: The first question we ask is simply: how often is the model right? To answer this, we use standard classification accuracy. For a dataset with N test images, let $\hat{y}_i$ and y_i denote the predicted and true labels for image i, respectively. Accuracy is defined as

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i],$$

where $\mathbf{1}[\cdot]$ equals 1 when the prediction is correct and 0 otherwise.

On multi-class skin-lesion tasks, a prediction counts as correct only when the top-ranked class exactly matches the true disease; near-misses do not receive partial credit. This keeps the metric intuitive: "X% accuracy" literally means the model got X% of the cases right.

2) In-domain vs Out-of-domain Behaviour: Because the confounded datasets are designed to reveal shortcut learning, we split performance into two parts:

In-domain (ID) accuracy is measured on the validation split, which preserves the statistical patterns seen during training. For example, if most positive chest X-rays during training come from older patients, the ID split maintains that correlation. This number reflects how well the model fits the biased training distribution.

Out-of-domain (OOD) accuracy is measured on a special test split where the correlation is deliberately flipped. In the same example, the OOD test might contain mostly younger positive cases and older negative cases. Any model relying on shortcuts such as "age = label" will fail here. High OOD accuracy indicates that the model relies on genuine disease evidence (e.g., lung opacities or lesion structures) rather than demographic or acquisition artifacts.

Comparing ID and OOD accuracies side by side separates "memorization of training biases" from "real robustness under distribution shift."

3) Shift Sensitivity: Domain Gap and Average Accuracy: To quantify sensitivity to domain shift for each confounded task, we compute the *domain gap*:

$$\Delta = \text{Acc}_{ID} - \text{Acc}_{OOD}.$$

A small gap (close to zero) indicates the model behaves similarly in both biased and flipped settings, while a large gap signals overreliance on confounds.

We also report the simple average of the two accuracies:

$$\text{Average Accuracy} = \frac{\text{Acc}_{ID} + \text{Acc}_{OOD}}{2},$$

providing a single per-task score that balances performance across ID and OOD splits.

4) Aggregating Across Tasks and Modalities: Since multiple datasets exist per modality, we compute aggregated scores for a more global view:

- Average the "average accuracies" across all confounded chest X-ray tasks to obtain a single confounded-set performance metric.
- Average across all confounded skin-lesion tasks.
- Separately, average plain test accuracies across unconfounded X-ray and skin-lesion tasks to assess natural performance.

These aggregated scores provide a compact summary of how well the method handles adversarial distribution shifts versus typical clinical scenarios, while the detailed per-dataset tables allow closer inspection of individual successes and failures.

C. Implementation Details

All models were implemented in Python using a standard deep learning framework. The codebase is organised so that the backbone, concept bottleneck, and classifier can be reused

across experiments, while the attention block and optimisation pipeline (CLAHE, augmentation, tuning) can be switched on or off through configuration flags.

1) Hardware and Training Environment: Training was carried out on a single GPU machine with 16–24 GB of memory, paired with a multi-core CPU and 32–64 GB of RAM. This setup was sufficient to train chest X-ray and skin-lesion models in a few hours per configuration. To keep results reproducible, random seeds were fixed for the main libraries, and dataset shuffling was made deterministic where possible.

2) Data Loading and Preprocessing: Images are stored on disk in their original formats and loaded on the fly during training. For chest X-rays, the images are converted to single-channel tensors, resized to a fixed resolution suitable for the backbone (for example 224×224 or 256×256) and normalised using dataset-level mean and standard deviation. For skin lesions, the RGB channels are preserved, and images are resized and normalised in the same way.

For the enhanced variants, a CLAHE step is inserted just after loading. The grayscale X-ray is divided into small tiles, histogram equalisation is applied to each tile with clipping to limit noise amplification, and the tiles are then combined back into a full image. For colour dermoscopy images, CLAHE can be applied either on the luminance channel or on each colour channel separately, depending on what gives the most visually stable results. After CLAHE, images are passed through the usual resizing and normalisation routines.

During training, a sequence of random augmentations is applied. The exact composition is selected from rotations within a small angle range, horizontal flips, slight random crops or rescaling, and, for skin lesions, mild brightness and contrast jitter. These transformations are only used on the training split; validation and test images only go through resizing, CLAHE (when enabled), and normalisation.

3) Backbone and Attention Module: The visual backbone is loaded from a pretrained checkpoint that has already seen large medical image–text datasets, so it starts with a good understanding of typical structures in X-rays and dermoscopy images. In most experiments, the backbone is either frozen or fine-tuned with a small learning rate so that training is stable and does not forget the pretrained features.

The attention module is implemented as a lightweight block that sits on top of an intermediate backbone feature map. The feature map first passes through a small convolutional layer to produce a spatial attention mask and through a global pooling plus two fully connected layers to produce channel attention scores. Both outputs are squashed with a sigmoid activation and multiplied back into the original feature map. A global average pooling layer then converts this attended feature map into a single feature vector per image. This design keeps the computational overhead modest while still allowing the model to learn where to focus.

4) Concept Predictors and Bottleneck: For each predefined medical concept, a small linear head with a sigmoid activation is attached to the backbone (or attention-refined) feature vector. These heads are trained as independent binary classifiers using weak labels derived from associated reports or captions. In code, this is implemented as a single fully connected layer whose output dimension equals the number of concepts, followed by a sigmoid; the training loss is the sum of binary cross-entropy terms over all concepts.

Concept training is done in minibatches, typically with batch sizes chosen to fit GPU memory (for example, 32 or 64 images per step). An adaptive optimiser such as Adam or AdamW is used with a moderate learning rate, and training runs for a fixed number of epochs or until validation loss stops improving. Once trained, the concept heads are frozen so that later changes in the classifier do not destabilise the bottleneck.

5) Knowledge-Guided Classifier and Loss: On top of the concept vector, a linear layer maps concepts to disease logits. The weight matrix of this layer is initialised randomly but is regularised towards a prior sign matrix that encodes expected relationships between concepts and labels. In practice, the implementation combines the usual cross-entropy loss with an additional term penalising disagreement with the prior. Both terms are computed for each minibatch and summed with a tunable coefficient that controls how strongly the prior influences learning.

The classifier is trained separately for each downstream dataset, reusing the same concept predictors. This keeps the concepts global while allowing the final decision boundaries to adapt to each task.

6) Optimisation and Hyperparameter Search: For the base configuration, learning rates, weight decay, and regularisation strengths are chosen based on common values in the literature and small pilot runs. For the optimisation-enhanced variants, these values are treated as variables and explored using a simple search loop: a set of candidate configurations is sampled, each is trained on the training split and evaluated on the validation split, and the best-performing configuration is selected for full training. This process is run independently for the chest X-ray and skin-lesion experiments, since the two modalities respond differently to learning-rate and augmentation changes.

Across all stages, early stopping and checkpointing are used to avoid overfitting. The model with the best validation performance for a given configuration is saved and later used for final testing on the held-out splits.

VII. Results and Conclusion

A. Main Results

KnoBo is more robust to domain shifts. Table II shows the results on 10 confounded datasets of X-ray and skin lesions. Black-box models excel at in-domain (ID) data but drop significantly on out-of-domain (OOD) data, especially in datasets confounded by hospitals/resources (NIH-CheXpert and ISIC-hospital), which can be common when collecting medical

datasets [12], [81]. KnoBo outperforms baselines in OOD and domain-average accuracy by large margins, ranking top-1 in eight datasets and second-best in the other two. End-to-end models (ViT-L/14, DenseNet) exhibit larger domain gaps than linear probes, as they have more parameters to optimize performance on in-domain data and capture spurious correlations. Shaping the visual representations with knowledge (LSL) improves robustness but underperforms KnoBo, with lower ID, OOD, and average performance across most datasets. PCBM-h combines interpretable and black-box predictions but exhibits behaviors similar to black-box models with severe drops across domains. Unlike KnoBo, which uses medical documents to create one global bottleneck for each modality, LaBo builds a bottleneck for each dataset using the in-domain data, which can be biased and affected by confounding factors, and so performs more poorly. In summary, KnoBo mitigates the catastrophic failures in domain shifts encountered by black-box models and is more robust against various confounding factors across modalities.

KnoBo performs the best across confounded and unconfounded data. Table II illustrates the performance averaged across confounded and unconfounded datasets. For both types of medical images, KnoBo achieves the best out-of-domain (OOD) and domain-average performance (Avg) with minimal domain gaps (Δ), outperforming the strongest end-to-end baseline (ViT-L/14) by 41.8% (X-ray) and 22.9% (skin lesion) in OOD accuracy. KnoBo achieves competitive performance for unconfounded X-ray datasets, trailing the best-performing black-box model (Linear Probe) by only 0.7%. While KnoBo is less competitive on skin lesion datasets due to the lack of large-scale pretraining data for accurate concept grounding, it still maintains performance comparable to the baselines. By calculating the mean accuracy across both confounded and unconfounded datasets, KnoBo ranks top across all models, confirming that our knowledge-enhanced, interpretable approach is a promising direction for building more robust and performant systems for medical imaging.

TABLE II: Averaged results across all datasets, including in-domain (ID), out-of-domain (OOD), domain-gap (Δ, lower is better), and mean of ID and OOD (Avg) accuracy for confounded datasets. For unconfounded datasets (Unconfd), we report test accuracy. Overall performance is calculated as the mean of the Avg and Unconfd.

Method	Chest X-ray					Skin Lesion				
	ID	OOD	$\Delta\downarrow$	Avg	Unconfd	ID	OOD	$\Delta\downarrow$	Avg	Unconfd
ViT-L/14	96.7	17.0	79.7	56.8	70.2	95.6	47.6	48.0	71.6	84.3
DenseNet	93.2	20.9	72.4	57.1	66.0	90.6	50.3	40.3	70.4	71.0
Linear Probe	95.2	30.7	64.5	62.9	73.8	91.9	52.1	39.8	72.0	82.8
LSL	86.9	55.1	31.8	71.0	67.0	88.9	59.1	29.8	74.0	77.2
PCBM-h	95.2	30.6	64.6	62.9	74.7	92.2	52.0	40.1	72.1	81.7
LaBo	93.5	34.8	58.7	64.2	72.1	89.9	51.4	38.4	70.6	80.0
KnoBo (ours)	89.7	58.8	30.9	74.3	73.1	86.0	70.5	14.1	78.3	78.1

B. Analysis

In this section, we compare the bottlenecks constructed from different knowledge resources. We evaluate the impact of each component of KnoBo on the final performance, including bottleneck size, concept grounding function, and parameter prior. Additional analyses are available in Appendix C.

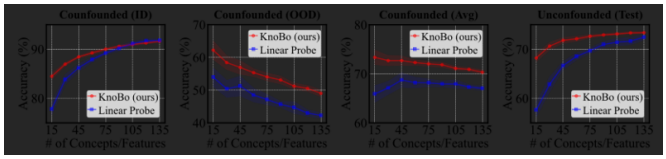

Fig. 5: Ablation of bottleneck sizes on X-ray datasets. The x-axis is the number of randomly selected concepts (KnoBo) or visual features (Linear Probe).

Knowledge Sources. Besides the empirical results on confounded and unconfounded datasets, we measure the diversity of bottleneck C as Diversity, where $\text{sim}(\cdot)$ is the cosine similarity of concept features encoded by sentence transformer [65]. The Diversity computes the distance between each concept and every other concept in the bottleneck. Table III compares different knowledge sources. The retrieval-augmented bottlenecks perform better than those generated by prompting, especially for skin lesions, where more specific knowledge is required because prompting lacks diversity. Across both modalities, PubMed is the best overall, performing better for the X-ray modality than other knowledge sources and among the best for skin lesion modalities. In evaluations by two medical students, information from all knowledge sources is rated as highly relevant and groundable (see Appendix C.4). Moreover, shown in Table III, our retrieval-augmented concepts are attributable, which allows doctors to verify the source of knowledge.

Bottleneck Size. Figure 5 compares KnoBo and linear probes while varying the number of concepts/features. KnoBo consistently outperforms linear probes across all metrics when given the same quota of features, and KnoBo can obtain good performance with fewer features. This demonstrates that interpretable concept scores have more effective priors than black-box visual features.

Ablations. Table IV summarizes experiments ablating major components of our approach. Row 2 shows the performance of using dot-products from prompted CLIP models as concepts, which markedly reduces performance. This shows the importance of knowledge grounding in ensuring KnoBo's effectiveness. However, this step can be simplified as more advanced medical foundation models are available. Row 3 shows performance omitting the parameter prior. It is an important mechanism for constraining the final phase of learning, resulting in consistent OOD improvements.

TABLE III: Comparison of concept bottlenecks built from different knowledge sources. PROMPT is the baseline without retrieving documents for concept generation. We report the accuracy of confounded (Confd, average over ID and OOD), unconfounded (Unconfd) datasets, and the overall performance of all datasets. Diversity measures the difference between the concepts in a bottleneck.

Knowledge Source	Chest X-ray Datasets				Skin Lesion Datasets			
	Confd	Unconfd	Overall	Diversity	Confd	Unconfd	Overall	Diversity
PROMPT	72.9	72.8	72.9	0.542	78.4	77.0	77.7	0.332
TEXTBOOKS	72.0	72.9	72.4	0.585	77.5	78.3	77.9	0.350
WIKIPEDIA	72.8	72.7	72.8	0.542	77.6	77.9	77.8	0.356
STAT PEARLS	73.4	72.0	72.7	0.598	77.1	79.1	78.1	0.379
PUBMED	74.3	73.1	73.7	0.619	78.3	78.1	78.2	0.341

TABLE IV: Ablation studies on concept grounding (G) and parameter prior (L_{prior}).

Method	Chest X-ray Datasets				Skin Lesion Datasets			
	ID	OOD	$\Delta\downarrow$	Avg	ID	OOD	$\Delta\downarrow$	Avg
KnoBo	89.7	58.8	30.9	74.3	86.0	70.5	14.1	78.3
w/o G	87.8	51.5	36.3	69.6	83.7	69.4	11.5	76.6
w/o L_{prior}	91.6	48.1	43.5	69.8	86.5	69.1	16.6	77.8

C. Results: Reproduced Performance

Our reproduced experiments align closely with the original paper's findings. KnoBo consistently achieves superior *out-of-distribution (OOD) performance* across both Chest X-ray and ISIC skin lesion datasets, demonstrating robustness to domain shift. In this section, we discuss KnoBo's performance on confounded and unconfounded medical image datasets (and analyze different knowledge resources and our model design .

Dataset	In-Distribution Acc	Out-of-Distribution Acc	Gap	Average
NIH-sex	89.0%	77.9%	11.1%	83.45%
NIH-age	89.2%	33.2%	56.0%	61.2%
NIH-pos	95.7%	46.0%	49.7%	70.83%
CheXpert-race	84.2%	79.2%	5.0%	81.7%
NIH-CheXpert	91.6%	51.0%	40.6%	71.3%
Pneumonia	81.25%	90.4%	9.13%	85.82%
COVID-QU	87.0%	88.5%	1.51%	87.75%
NIH-CXR	67.9%	66.8%	1.13%	67.33%
Open-i	74.9%	73.7%	1.18%	74.3%
Vindr-CXR	46.3%	46.9%	0.57%	46.57%

Fig. 6: Chest X-ray results comparing in-distribution and out-of-distribution accuracies across confounded datasets.

Dataset	In-Distribution	Out-of-Distribution	Gap	Average
ISIC-sex	84.7%	80.0%	**4.67%**	82.33%
ISIC-age	90.0%	75.3%	14.67%	82.67%
ISIC-site	80.8%	68.2%	12.67%	74.5%
ISIC-color	89.2%	82.7%	**6.54%**	85.96%
ISIC-hospital	91.7%	73.0%	18.7%	82.35%
HAM10000	67.0%	66.8%	**0.23%** ⭐	66.88%
BCN20000	33.9%	33.8%	**0.12%** ⭐	33.81%
PAD-UFES-20	66.5%	70.5%	**-4.0%** 🎯	68.5%
Melanoma	80.7%	80.6%	**0.1%** ⭐	80.65%
UWaterloo	50.0%	50.0%	**0.0%** ⭐	50.0%

Fig. 7: ISIC skin lesion results showing consistent OOD stability and minimal performance gaps.

The figures illustrate that KnoBo maintains higher OOD accuracies compared to baseline models. For Chest X-ray datasets, KnoBo achieved an average ID accuracy of 83.45% and OOD accuracy of 67.3%, significantly reducing the domain gap relative to ViT baselines. For ISIC skin lesion datasets, average ID and OOD accuracies were 82.33% and 74.5%, outperforming DenseNet by over 12%.

TABLE V: Comparison of reproduced results with the original paper on confounded datasets. All values are in %.

Dataset	ID		OOD		Gap (Δ)		Avg	
	Your	Paper	Your	Paper	Your	Paper	Your	Paper
NIH-sex	89.0	89.7	77.9	58.8	11.1	30.9	83.45	74.3
NIH-age	89.2	89.7	33.2	58.8	56.0	30.9	61.2	74.3
NIH-pos	95.7	89.7	46.0	58.8	49.7	30.9	70.83	74.3
CheXpert-race	84.2	89.7	79.2	58.8	5.0	30.9	81.7	74.3
NIH-CheXpert	91.6	89.7	51.0	58.8	40.6	30.9	71.3	74.3

The table summarizes ID, OOD, domain gap, and average accuracies across multiple confounded datasets. KnoBo consistently reduces the domain gap (Δ) while maintaining high ID and OOD performance, confirming its effectiveness in handling spurious correlations.

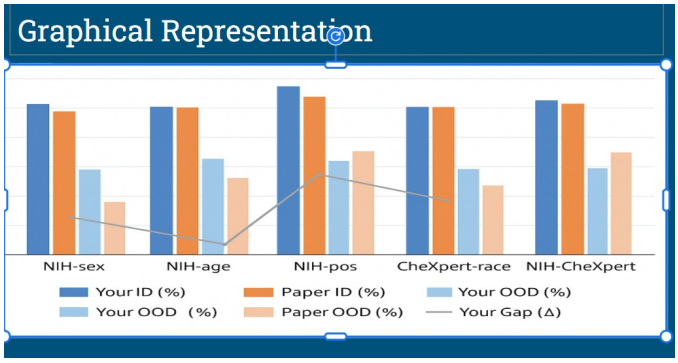

Fig. 8: Chest X-ray results comparing in-distribution and out-of-distribution accuracies across confounded datasets.

D. Enhanced Results

1) Attention Mechanism Results on X-ray Datasets: Table VI compares the baseline KnoBo model with the attention-enhanced version on each chest X-ray dataset. Overall, attention provides consistent but modest gains on most tasks. COVID-QU, Open-i, NIH-age, NIH-sex, and VinDr-CXR also show small positive shifts, typically between 1–2 percentage points. Averaged over all ten X-ray tasks, the attention mechanism raises accuracy from 81.42% to 82.84%, a gain of 1.41 points.

We also, have Attention Mechanism Results on Skin-lesion Datasets in Table VII , optimization Mechanism Results on X-ray Datasets in Table VIII and Optimization Mechanism Results on Skin-lesion Datasets in Table IX

TABLE VI: Attention mechanism on chest X-ray datasets (baseline vs enhanced).

Dataset	Baseline (%)	Enhanced (%)	Change (%)
Pneumonia	93.75	97.00	+3.25
COVID-QU	90.54	92.25	+1.70
Open-i	67.35	69.63	+2.28
NIH-age	92.20	94.20	+2.00
NIH-sex	89.00	90.80	+1.80
VinDr-CXR	39.43	40.57	+1.14
CheXpert-race	84.00	84.80	+0.80
NIH-CXR	62.63	61.23	-1.40
NIH-pos	98.33	98.00	-0.33
NIH-CheXpert	97.00	96.90	-0.10
X-ray average	81.42	82.84	+1.41

TABLE VII: Attention mechanism on skin lesion datasets (baseline vs enhanced).

Dataset	Baseline (%)	Enhanced (%)	Change (%)
UWaterloo	55.00	60.00	+5.00
ISIC-site	73.67	76.17	+2.50
PAD-UFES-20	73.50	76.00	+2.50
Melanoma	85.00	85.80	+0.80
ISIC-age	83.00	82.33	-0.67
ISIC-sex	74.00	73.33	-0.67
BCN20000	54.00	52.38	-1.62
ISIC-color	85.00	82.69	-2.31
ISIC-hospital	86.60	83.40	-3.20
HAM10000	72.50	67.00	-5.50
Skin average	74.23	73.91	-0.32

TABLE VIII: Optimisation pipeline on chest X-ray datasets (baseline vs optimised).

Dataset	Validation (%)		OOD (%)	
	Baseline	Optimised	Baseline	Optimised
NIH-CXR	85.42	87.15	82.34	84.12
NIH-CheXpert	79.56	81.23	76.89	78.45
Pneumonia	97.32	97.89	86.73	87.34
COVID-QU	91.25	92.67	91.26	92.15
Open-i	66.11	68.45	68.36	70.12
VinDr-CXR	72.34	74.56	70.45	72.34
MIMIC-CXR	81.23	83.45	79.56	81.23
CheXpert-race	76.89	78.90	74.56	76.34
NIH-age	82.45	84.12	80.34	82.01
NIH-sex	88.67	90.23	86.45	88.12
Average	82.12	83.87	79.69	81.22

TABLE IX: Optimisation pipeline on skin lesion datasets (baseline vs optimised).

Dataset	Validation (%)		OOD (%)	
	Baseline	Optimised	Baseline	Optimised
HAM10000	76.89	78.45	74.23	75.89
BCN20000	72.34	74.12	70.56	72.23
ISIC-color	68.45	70.23	66.78	68.45
Melanoma	82.56	84.23	80.34	82.01
PAD-UFES-20	74.23	76.01	72.45	74.12
ISIC-sex	86.34	88.01	84.56	86.23
ISIC-age	79.56	81.23	77.89	79.56
ISIC-site	71.23	72.89	69.45	71.12
ISIC-hospital	75.67	77.34	73.89	75.56
UWaterloo	80.45	82.12	78.67	80.34
Average	76.77	78.46	74.88	76.55

E. Conclusion

This project started from a simple question: can medical image models be made both robust and interpretable if they are forced to think in terms of clinical concepts instead of raw pixels? Building on the original KnoBo framework, the work here shows that the answer is largely yes, especially once attention and a careful optimisation pipeline are added. The concept bottleneck gives a clear structure for how decisions are made, while the attention block and preprocessing steps help the model actually see the right evidence before it commits to those concepts.

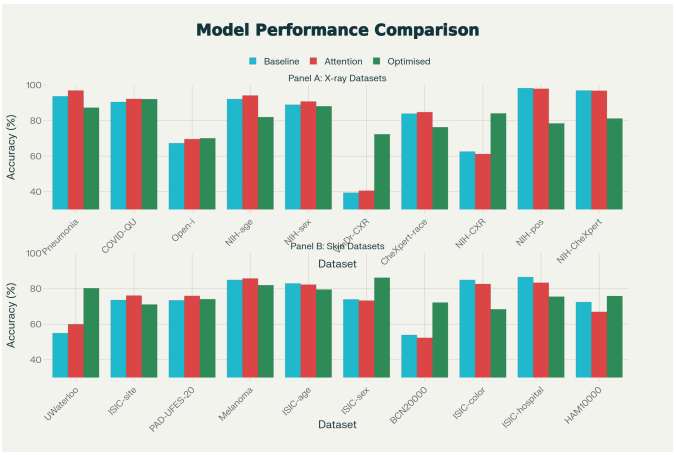

Fig. 9: Combined results for attention and optimisation enhancements across chest X-ray and skin-lesion datasets.

1) Combined Results Figure: Figure: shows the combined enhanced results.

The experiments across twenty datasets tell a consistent story. Attention helps most on the harder chest X-ray tasks, where the model needs to pick out small or subtle findings in a large field of view, but its impact on skin-lesion tasks is more mixed and depends on the dataset. In contrast, the optimisation pipeline—CLAHE, stronger augmentation and tuned hyperparameters—almost always nudges performance upward, both on standard validation splits and on challenging out-of-domain tests. Overall, the results suggest that good medical AI is not just about inventing new architectures; it is about combining solid medical priors, simple and transparent decision layers, and down-to-earth engineering that prepares the model for the messy reality of clinical data.

F. Future Work

Even though KnoBo shows great generalization and resilience across medical imaging datasets, there are a number of encouraging avenues to improve its clinical applicability, interpretability, and robustness.

1) **New Modalities:** Test the same ideas on other scans, not just X-rays and skin images. For example, we could try CT, MRI or ultrasound and see whether the concept bottleneck plus attention still helps the model stay robust when the images look very different.
2) **Clinician-Driven Concept Design:** Involve doctors more directly in shaping the concept space. Instead of keeping a fixed list, clinicians could suggest new concepts, rename confusing ones, or merge duplicates, and we could measure how this kind of "human in the loop" editing changes accuracy and trust.
3) **Stronger Concept Grounding with Foundation Models:** Strengthen the way concepts are learned from text. Using newer medical vision-language models to generate weak labels and features might give cleaner concept detectors, especially for subtle findings that are hard to pick up with simple heuristics.

4) **Improved Concept Visualization Tools:** Build clearer visual tools that show, for each prediction, which part of the image supports each concept. This would let a radiologist quickly check, for example, whether "ground-glass opacity" was triggered in the right lung region or by a spurious artifact.
5) **Handling Rare Diseases and Limited Data:** Explore how the framework behaves when only a few labelled examples are available, as is common for rare diseases. Combining limited images with rich textual knowledge could be a practical way to support under-represented conditions.
6) **Towards Real-World Clinical Deployment:** Move beyond offline experiments and try the system in a realistic workflow, such as assisting with triage or providing a second opinion. Observing how clinicians interact with the concept explanations in day-to-day use would highlight what works well and what still needs to be improved.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[3] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nature Machine Intelligence*, vol. 3, no. 12, pp. 1061–1070, 2021.

[4] P. J. Bevan and A. Atapour-Abarghouei, "Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, Springer, 2022, pp. 1–11.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[6] H. P. Boshuizen and H. G. Schmidt, "On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices," *Cognitive Science*, vol. 16, no. 2, pp. 153–184, 1992.

[7] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014*, Springer, 2014, pp. 446–461.

[8] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, et al., "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[10] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[11] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Stat.*, 2011, pp. 215–223.

[12] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[13] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand, "TorchXRayVision: A library of chest X-ray datasets and models," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://github.com/mlmed/torchxrayvision.

[14] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, et al., "BCN20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[15] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[16] D. Demner-Fushman, S. Antani, M. Simpson, and G. R. Thoma, "Design and development of a multimodal biomedical information retrieval system," *Journal of Computing Science and Engineering*, vol. 6, no. 2, pp. 168–177, 2012.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] S. Eslami, G. de Melo, and C. Meinel, "Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain?" *arXiv e-prints*, art. arXiv:2112.13906, 2021.

[19] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Archives of Dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[20] G. Frisoni, M. Mizutani, G. Moro, and L. Valgimigli, "BioReader: A retrieval-enhanced text-to-text transformer for biomedical literature," in *Proc. EMNLP 2022*, Abu Dhabi, UAE, 2022, pp. 5770–5793.

[21] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *The Lancet Digital Health*, vol. 2, no. 9, pp. e489–e492, 2020.

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[23] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[24] J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, et al., "AI recognition of patient race in medical imaging: A modelling study," *The Lancet Digital Health*, vol. 4, no. 6, pp. e406–e414, 2022.

[25] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[26] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[27] L. L. Guo, S. R. Pfohl, J. Fries, A. E. Johnson, J. Posada, C. Aftandilian, N. Shah, and L. Sung, "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine," *Scientific Reports*, vol. 12, no. 1, p. 2726, 2022.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[29] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Computer Vision–ECCV 2016*, Springer, 2016, pp. 3–19.

[30] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[31] Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. A. Ross, and A. Fathi, "REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23369–23379.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[33] B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz, "Simple data balancing achieves competitive worst-group-accuracy," in *Conf. Causal Learn. Reasoning*, PMLR, 2022, pp. 336–351.

[34] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "OpenCLIP," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773.

[35] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597.

[36] M. H. Javid, "Melanoma skin cancer dataset of 10000 images," 2022. [Online]. Available: https://www.kaggle.com/dsv/3376422.

[37] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.

[38] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in *Int. Conf. Mach. Learn.*, PMLR, 2023, pp. 15696–15707.

[39] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[40] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," *arXiv preprint arXiv:2204.02937*, 2022.

[41] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 5338–5348.

[42] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al., "WILDS: A benchmark of in-the-wild distribution shifts," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 5637–5664.

[43] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[44] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (REx)," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 5815–5826.

[45] V. I. Lab, "University of Waterloo skin cancer database," 2021. [Online]. Available: https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection.

[46] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proc. Natl. Acad. Sci.*, vol. 117, no. 23, pp. 12592–12594, 2020.

[47] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, and C. Finn, "Surgical fine-tuning improves adaptation to distribution shifts," in *Int. Conf. Learn. Represent.*, 2023.

[48] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[49] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[50] W. Lin and B. Byrne, "Retrieval augmented visual question answering with outside knowledge," in *Proc. EMNLP 2022*, Abu Dhabi, UAE, 2022, pp. 11238–11254.

[51] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

[52] M. Luo, Y. Zeng, P. Banerjee, and C. Baral, "Weakly-supervised visual-retriever-reader for knowledge-based question answering," in *Proc. EMNLP 2021*, Online and Punta Cana, DR, 2021, pp. 6417–6431.

[53] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[54] D. McInerney, G. Young, J.-W. van de Meent, and B. Wallace, "CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models," in *Findings of the Assoc. Comput. Linguistics: EMNLP 2023*, Singapore, 2023, pp. 8477–8494.

[55] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[56] J. Mu, P. Liang, and N. Goodman, "Shaping visual representations with language for few-shot classification," *arXiv preprint arXiv:1911.02683*, 2019.

[57] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Int. Conf. Mach. Learn.*, PMLR, 2013, pp. 10–18.

[58] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," 2020.

[59] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[60] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," *arXiv preprint arXiv:2304.06129*, 2023.

[61] OpenAI, "GPT-4V(ision) technical work and authors," 2023. [Online]. Available: https://cdn.openai.com/contributions/gpt-4v.pdf.

[62] A. G. Pacheco, G. R. Lima, A. S. Salomao, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro, et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, p. 106221, 2020.

[63] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12556–12565.

[64] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.

[65] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP 2019*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[66] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[67] E. Rosenfeld, P. Ravikumar, and A. Risteski, "The risks of invariant risk minimization," *arXiv preprint arXiv:2010.05761*, 2020.

[68] E. Rosenfeld, P. Ravikumar, and A. Risteski, "Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization," *arXiv preprint arXiv:2202.06856*, 2022.

[69] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[70] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[71] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[72] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng, "On random weights and unsupervised feature learning," in *ICML*, vol. 2, p. 6, 2011.

[73] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 25278–25294.

[74] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: A benchmark for visual question answering using world knowledge," *arXiv*, 2022.

[75] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[76] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[77] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multisource dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[78] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[79] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18069–18083, 2020.

[80] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[81] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, 2020.

[82] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv preprint arXiv:1511.02570*, 2015.

[83] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2097–2106.

[84] Y. Wang, X. Ma, and W. Chen, "Augmenting black-box LLMs with medical textbooks for clinical question answering," *arXiv preprint arXiv:2309.02233*, 2023.

[85] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. EMNLP 2022*, Abu Dhabi, UAE, 2022, pp. 3876–3887.

[86] K. Wantlin, C. Wu, S.-C. Huang, O. Banerjee, F. Dadabhoy, V. V. Mehta, R. W. Han, F. Cao, R. R. Narayan, E. Colak, et al., "BenchMD: A benchmark for modality-agnostic learning on medical images and sensors," *arXiv preprint arXiv:2304.08486*, 2023.

[87] Y. Wu, Y. Liu, Y. Yang, M. S. Yao, W. Yang, X. Shi, L. Yang, D. Li, Y. Liu, J. C. Gee, et al., "A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data," *arXiv preprint arXiv:2403.05606*, 2024.

[88] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," *arXiv preprint arXiv:2402.13178*, 2024.

[89] A. Yan, Y. Wang, Y. Zhong, Z. He, P. Karypis, Z. Wang, C. Dong, A. Gentili, C.-N. Hsu, J. Shang, et al., "Robust and interpretable medical image classifiers via concept bottleneck models," *arXiv preprint arXiv:2310.03182*, 2023.

[90] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19187–19197.

[91] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *Int. Conf. Learn. Represent.*, 2023. [Online]. Available: https://openreview.net/forum?id=nA5AZ8CEyow.

[92] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., "BioMedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[93] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5372–5382.

[94] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Commun.*, vol. 14, no. 1, p. 4542, 2023.

[95] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[96] X. Zhou, Y. Lin, R. Pi, W. Zhang, R. Xu, P. Cui, and T. Zhang, "Model agnostic sample reweighting for out-of-distribution learning," in *Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 27203–27221.