Concept Complement Bottleneck Model for Interpretable Medical Image Diagnosis

Hongmei Wang HWANGFY@CONNECT.UST.HK
Junlin Hou CSEJLHOU@UST.HK
Sunan He SHEBD@CONNECT.UST.HK
Shu Yang SYANGCW@CONNECT.UST.HK
Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

Hao Chen ∗ SYANGCW@CONNECT.UST.HK
Department of Computer Science and Engineering, Department of Chemical and Biological Engineering, Division of Life Science, and State Key Laboratory of Nervous System Disorders, Hong Kong University of Science and Technology, Hong Kong, China.
HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China.

Editors: Under Review for MIDL 2026

Abstract

Models based on human-understandable concepts have received extensive attention to improve model interpretability for trustworthy artificial intelligence in the field of medical image analysis. These methods can provide convincing explanations for model decisions but heavily rely on detailed annotations of predefined concepts. Consequently, they may be ineffective when concepts or annotations are incomplete or of low quality. Although some methods can automatically discover novel and effective visual concepts instead of relying on predefined ones, or generate human-understandable concepts using large language models, they often deviate from medical diagnostic evidence and remain difficult to interpret. In this paper, we propose a concept complement bottleneck model for interpretable medical image diagnosis. Specifically, we use cross-attention modules to extract key image features related to the predefined textual concepts and employ independent concept adapters and bottleneck layers to distinguish concepts more effectively. Additionally, we devise a concept complement module to mine local concepts from the concept bank constructed using medical literature. The model jointly learns human-annotated predefined concepts and automatically discovered ones to improve performance in concept detection and disease diagnosis. Comprehensive experiments demonstrate that our model outperforms state-of-the-art methods while providing diverse and interpretable explanations.

Keywords: Explainable AI, medical image diagnosis, concept bottleneck model, concept discovery.

1. Introduction

Despite the impressive performance in Medical Image Analysis (MIA), AI models still face several challenges in clinical deployment. A critical challenge is that such black-box models lack transparency and interpretability throughout end-to-end training. Within medical image analysis, clinicians and patients need to understand how data are represented in the model's feature space and the evidence for model decisions. Such understanding fosters trust in the model's predictions and ultimately supports the deployment of AI in clinical

practice (Lucieri et al., 2020). Concept-based post-hoc models have attracted the most attention among explainable AI approaches because they use human-understandable textual or human-friendly visual concepts to explain model decisions (Gupta and Narayanan, 2024). Some studies predict concepts that are present in images and have been densely annotated by clinicians to drive subsequent decisions. The Concept Bottleneck Model (CBM) (Koh et al., 2020) is a representative example. CBM first learns concept scores by minimizing the cross-entropy loss of concept detection. Variants of CBM have been widely explored for disease diagnosis using X-ray, ultrasound, and other imaging modalities (Chauhan et al., 2023; Marcinkevičs et al., 2024). Although CBM-based methods improve interpretability, they demand complete concept annotations for the entire training set. In clinical practice, obtaining such dense labels is extremely time-consuming and labor-intensive. Some approaches automatically discover visual concepts (Fang et al., 2020; Kong et al., 2024), yet these concepts rarely align with clinical terminology in a general domain. Leveraging Large Language Models (LLMs) or Vision-Language Models (VLMs), recent work automatically generates textual concepts for images (Shang et al., 2024; Yang et al., 2023). Nevertheless, such concepts only remain aligned with general-domain knowledge and often lack clinical relevance. Although these methods deliver competitive interpretability and accuracy, they either depend entirely on human-annotated concepts or rely solely on automatically discovered ones for decision-making. The former demands dense concept labels yet still underperforms black-box models, whereas the latter often drifts away from authentic clinical reasoning. Moreover, existing concept-based approaches typically extract a single image feature map and reuse it for all concepts, ignoring inter-concept heterogeneity. Visually simple concepts are easier to extract yet usually carry limited discriminative power, particularly for hard-to-diagnose cases.

To tackle these limitations, we introduce the Concept Complement Bottleneck Model (CCBM). CCBM automatically mines local textual concepts from medical literature to enrich the global predefined concept set, yielding higher accuracy and richer explanations without demanding extra manual labels. The CCBM framework introduces a clinically-aligned concept bank derived from medical textbooks to ensure relevance and validity. It employs per-concept processing with channel-wise attention, allowing each concept to have its own adapter and bottleneck layer to capture specific visual evidence effectively. Additionally, a concept complement module dynamically merges local concepts with a global set to enhance interpretability. Empirically, CCBM achieves significant performance improvements in disease diagnosis and concept detection on multiple medical imaging datasets, with verified interpretability through quantitative and qualitative analyses.

2. Methodology

In this section, we will introduce how to construct concept banks and the detailed structure of CCBM. The framework of our method is shown in Fig. 1.

2.1. Global Predefined Textual Concept Learning

The global predefined textual concepts are the human-annotated concepts in datasets. For predefined concept learning, we encode the textual concepts and get the corresponding

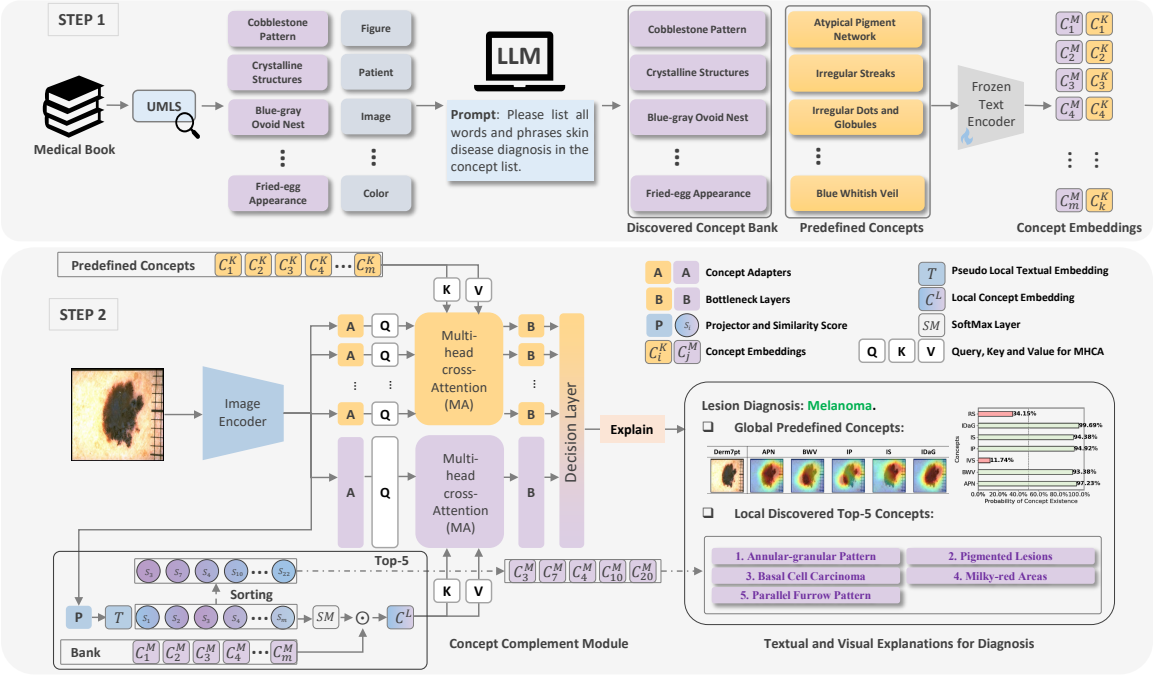


Figure 1: **Concept Complement Bottleneck Model**. STEP 1: Concept Bank Construction. STEP 2: Concept Complement Bottleneck Model. The CCBM contains two multi-head attention branches to encode image features corresponding to each concept to guide the model decision. The first is to learn the visual features of global predefined concepts and the second is to discover effective local concepts in the textual concept bank via the concept complement module.

visual embedding of these concepts from an image encoder to provide inputs for the Multi-head cross-Attention (MA) module. The textual known concept set is defined as $\mathcal{C}^{\mathcal{K}}$ which includes k predefined labeled concepts, and the textual known concept encoder E_T is used to encode the textual known concepts to the subspace with d_c dimensions: $E_T : \mathcal{C}^{\mathcal{K}} \rightarrow \mathbb{R}^{d_c}$. Furthermore, the concept embeddings C_i^K $(i = 1, 2, ..., k)$ will be used as the keys and values in the MA module for predefined concept learning.

2.2. Local Textual Concept Discovery

2.2.1. Concept Bank Construction.

For local concept discovery, we collect human-friendly textual medical concepts to build medical concept banks for different medical diagnosis tasks first. As STEP 1 shown in Fig.1, for each task, we first extract concepts by Unified Medical Language System (UMLS) (Bodenreider, 2004) from medical books that are highly related to the specific task. However, not all extracted concepts are related to medicine, such as "Figure", "Image", "Criteria", etc. Therefore, we feed all extracted concepts into an LLM (GPT-4o) and give a suitable prompt to retain high-related medical concepts and remove low-related concepts. In this

way, we build medical concept banks for diagnosis tasks easily. We formulate the concept bank with m concepts as $\mathcal{C}^{\mathcal{M}}$. Similarly, using the same text encoder mentioned in the section 2.1, we can get the textual embeddings C_i^M $(i = 1, 2, ..., m)$ for medical concepts.

2.2.2. Concept Complement Module.

The concept complement module aims to project image features into the textual concept embedding subspace and learns local concepts from the concept bank. In this module, a projector P projects basic image feature $\mathbf{f}$ to the textual concept embedding subspace and generates the pseudo local textual concept embedding T. The embedding T will be used to calculate the cosine similarity with all textual concept embeddings in the concept bank. Next, these similarities will be activated by the $softmax$ layer to generate the possibility vector to weigh all textual concept embeddings in the concept bank. For each image, the textual concept with the largest weight will be discovered as one of the local textual concepts for the input image. We formulate the concept complement module:

$$s_j = (C_j^M \cdot T)/(\|C_j^M\|\|T\|), T = P(\mathbf{f}), \tag{1}$$

$$\mathbf{w} = softmax([s_1, s_2, ..., s_m]), \tag{2}$$

where C_j^M is the j-th concept embeddings in concept bank and s_j means the cosine similarity of C_j^M. Then, we can obtain the local textual concept embedding for each image as $C^L = \mathbf{w} \cdot [C_1^M, C_2^M, ...C_m^M]$. For each image, based on the values of $\mathbf{w}$, we can get the importance scores of each concept in the concept bank.

2.3. Concept Complement Bottleneck Model

2.3.1. Dual Multi-Head Attention Branches for Concept Learning.

Based on the branches of global predefined concept learning and local textual concept discovery, we can build the CCBM using the following steps. E_T is the frozen text encoder to encode the textual concepts in the predefined concept set and concept bank. E_I is the image encoder to map the input images $\mathcal{X}$ including n samples from n_c classes to the image feature space: $E_I : \mathcal{X} \to \mathbb{R}^d$, where d represents the feature dimension. Intuitively, if one feature is used to calculate cross-attention effectively with all visual concept embeddings, the image encoder needs to be strong enough to extract all crucial visual features for all different textual concepts. However, difficult concepts are hard to capture, so we use concept adapters to capture the differences in visual concepts better. Hence, k concept adapters $A_i^K (i = 1, ..., k)$ are set to extract specific concept features from the basic image features to be the queries Q of the predefined concept MA module. Similarly, concept adapters A^L extract visual features of the local concepts. In our setting, each concept adapter is set as a fully connected layer that maps the d-dimension feature space to the d_c-dimension subspace:

$$Q_i^K = A_i^K(E_I(\mathcal{X}))(i = 1, 2, ..., k), Q^L = A^L(E_I(\mathcal{X})) \tag{3}$$

where Q_i^K is the i-th query in the MA module for predefined concept learning and Q^L is the query in the MA module for local concept learning. The predefined concept embedding $\{C_1^K, C_2^K, ..., C_k^K\} = E_T(C^K)$ will be used as the keys and values in the MA module for

predefined concept learning, and the local concept embedding C^L will be used as the key and value in the MA module for local concept learning. We only formulate the single-head cross-attention mechanism instead of complicated MA formulation:

$$Attn(Q,K) = softmax\left((QK^T)/(\sqrt{d_c})\right), Attn_w(Q,K,V) = Attn(Q,K)V \tag{4}$$

where $Attn$ is the attention map matrix whose elements represent the weights of different concept pairs and $Attn_w$ is the weighted sum of all concept embeddings. Furthermore, we calculate the concept scores based on these attention scores. Different from the previous bottleneck models (Koh et al., 2020), which directly use one bottleneck layer to get all concept scores, we use an independent bottleneck layer to get the score for each concept.

2.4. Concept Complement Bottleneck Model for Medical Image Diagnosis

2.4.1. Explainable Diagnosis Decision using Predefined Concept Set

If we do not set the local concept learning branch in our model, based on these concept scores, we can directly predict the diagnosis results by a decision layer f_d:

$$\hat{Y} = f_d(\mathbf{g}) \in \mathbb{R}^c, \tag{5}$$

where $\mathbf{g} = [g_1, g_2, ..., g_k]$ is the concept score vectors of input images and $\hat{Y}$ is the final disease prediction. During the training process, we jointly train the model to perform well on the concept detection task and disease diagnosis task. In particular, we require model decisions to more explicitly depend on these concept scores to ensure model interpretability. Therefore, we leverage the cross-entropy loss for the classification task and the concept-learning loss for the concept detection task. The total loss of our CCBM is:

$$\min_{\hat{Y},G}\left(\lambda_1 \mathcal{L}_{ce}(\hat{Y},Y) + \mathcal{L}_{cep}(G,C)\right), \tag{6}$$

where $\hat{Y}$ is the classification prediction of the model, Y is the ground truth of image diagnosis, C is the matrix of the ground truth of the concept detection task, and λ_1 is the hyperparameter to balance the two tasks. The cross-entropy loss for classification task and the multi-label classification cross-entropy loss for concept learning are: $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}), \tag{7}$$

$$\mathcal{L}_{cep} = -\sum_{i=1}^{N}\sum_{j=1}^{n_k}\left(c_{ij}\log(g'_{ij}) + (1-c_{ij})\log(1-g'_{ij})\right), \tag{8}$$

2.4.2. Concept Complement Strategy

Although existing models based on concept bottleneck can maintain high classification performance while providing interpretability by training on concept detection tasks and classification tasks, the model performance needs to be balanced between concept detection

and classification tasks. In addition, there is still a gap between explainable models and black-box models in classification tasks. In order to reduce this performance gap while maintaining model transparency, we propose a concept complement strategy to learning unknown concepts that are helpful for diagnosis. For our concept complement strategy, based on the concept scores, the CCBM makes medical decisions using a linear decision layer. The final prediction is $\hat{Y} = f_d(\mathbf{g}) \in \mathbb{R}^{n_c}$ where $\mathbf{g} = [g_1, g_2, ..., g_k, g_{k+1}]$ is the concept score vector of the input image, and the first k elements are the predefined concept scores and the last element indicates the total score of discovered local concepts. During the training process, we leverage the cross-entropy loss $\mathcal{L}_{ce}$ for the classification task and the multi-label binary cross-entropy loss $\mathcal{L}_{bce}$ for the concept detection task. The loss of CCBM is $(\lambda\mathcal{L}_{ce} + \mathcal{L}_{bce})$ where λ is used to balance the two tasks.

3. Experiments

3.1. Experiment Settings

We conducted comprehensive experiments on two dermoscopic image datasets *Derm7pt* and *Skincon*, and an ultrasound breast image dataset *BrEaST*. **Derm7pt** (Kawahara et al., 2018) contains 1011 dermoscopic images, including 20 specific skin disease diagnoses and detailed labels of 7 clinical concepts based on the seven-point skin lesion malignancy checklist. We only considered 827 images in which the diagnosis belongs to *Nevus* and *Melanoma*. **Skincon** (Daneshjou et al., 2022) contains 3230 skin images of *Malignant*, *Benign*, and *Non-neoplastic* categories in the Fitzpatrick 17k dataset (Groh et al., 2021). We chose 22 concepts where there are at least 50 images containing the concept from 48 general medical clinical concepts densely annotated by two dermatologists. **BrEaST** (Pawłowska et al., 2024) is an ultrasound breast image dataset with detailed annotations via 7 concepts from BI-RADS descriptors, which contains 256 images with 3 different types of breast diagnosis, including *Benign*, *Malignant* and *Normal*. We only use 254 abnormal breast images in our experiments. The used concepts are detailedly in Table 1. As for dermoscopic image datasets *Derm7pt* and *Skincon*, we follow PCBM (Mert Yuksekgonul et al., 2022) and choose the trained Inception-v3 model (Szegedy et al., 2015) as the image encoder to ensure fairness since the PCBM used the backbone, and we use pretrained ResNet50 as the image encoder for other two datasets. The frozen pre-trained ClinicalBERT (Alsentzer et al., 2019) is utilized as the text encoder. In our experiments, the concept adapters and aggregators are set as FCLs. Additionally, we use Adam optimizer for model training. We use the grid method to choose model hyperparameter λ. To evaluate model performance, we use the Area Under Curve (AUC), ACCuracy (ACC) and F1-score as disease diagnosis evaluation metrics, and the first two metrics are also used to evaluate the concept detection tasks of *Derm7pt*, *Skincon* and *BrEaST*. All of mean and standard deviation results are obtained by five-fold cross-validation in our experiments.

3.2. Comparison Algorithms

To verify the effectiveness and advancement of the model, we compare CCBM with the state-of-the-art methods, including CBM (Koh et al., 2020), PCBM (-H) (Mert Yuksekgonul et al., 2022), an Ante-hoc Explainable Concept-based model (AEC) (Sarkar et al., 2022),

Table 1: Dataset Details in Our Experiments.

Dataset	Used Concepts
Derm7pt	Atypical Pigment Network (APN), Blue Whitish Veil (BWV), Irregular Vascular Structures (IVS), Irregular Pigmentation (IP), Irregular Streaks (IS), Irregular Dots and Globules (IDaD), Regression Structures (RS)
Skincon	PAPule (PAP), PLAque (PLA), PUStule (PUS), BULla (BUL), PATch (PAT), NODule (NOD), ULCer (ULC), CRUst (CRU), EROsion (ERO), ATRophy (ATR), EXUdate (EXU), TELangiectasia (TEL), SCALe (SCAL), SCAR (SCAR), FRIable (FRI), Dome-SHaped (DSH), Brown-Hyperpigmentation (BrH), White-Hypopigmentation (WhH), PURple (PUR), YELlow (YEL), BLAck (BLA) and ERYthema (ERY)
BrEaST	Irregular SHape (IRS), Not Circumscribed Margin (NCM), Hyperechoic or Heterogeneous Echogenicity (HoHE), Posterior Features (PF), Hyperechoic Halo (HH), CALcifications (CAL), Skin Thickening (ST)

Concept-Based Interpretability using Vision-Language Models (CBIVLM) (Patrício et al., 2023) and Energy-based CBM (ECBM) (Xu et al., 2024). We also test the backbone models to evaluate the gap between our explainable model and black-box models. For methods that do not support concept detection, we use "N/A" to indicate invalid data.

3.3. Experimental Results and Analysis

3.3.1. Disease Diagnosis and Concept Detection

To verify the effectiveness of our model on the concept detection task and disease diagnosis task, we conducted extensive experiments and compared CCBM with the state-of-the-art methods on three datasets using multiple metrics. The five-fold cross-validation results of CCBM on two tasks are shown in Table 2. CCBM achieves outstanding performance for dermoscopic image analysis. For the *Derm7pt* dataset, CCBM achieves the best performance in the concept detection task, significantly outperforming the comparison explainable methods with 93.15% AUC, the 87.18% ACC and 84.20% F1-score. Additionally, CCBM surpasses the black-box model on all classification task metrics. Regarding the *Skincon* dataset, CCBM excels in the concept detection task on the AUC 85.51%, which is higher than other competitors, and the best ACC of 91.32%. While maintaining strong concept detection performance, the AUC reaches 85.15%, outperforming other competitors, and the ACC achieves an impressive 80.22%, which is comparable with the best model, AEC. The black-box model underperforms on the dataset, with CCBM surpassing it by 5% on AUC, 3% on AUC and 7% on F1-score. For ultrasound diagnosis and analysis, the results of the *BrEaST* dataset show that CCBM achieves the best performance for ultrasound image analysis on the concept detection task and the classification task on all evaluation metrics. The concept detection AUC is 76.22% which is 6% higher than CBM. The classification AUC, ACC and F1-score are 88.89%, 78.81% and 78.81%, respectively, representing a significant improvement. In comparison to the black-box model, CCBM outperforms ResNet50 across all metrics, while CBM is comparable to the black box model. Overall, CCBM learns the unknown concepts to complement the predefined known concepts to provide additional information for the model decision, helping to outperform other advanced explainable models

and the compared black-box model on disease diagnosis while guaranteeing the best performance on concept detection. Detailed concept detection results and the discovered concepts are provided in Appendix.A and Appendix.B.

Table 2: Quantitative results on disease diagnosis and concept detection tasks with comparison methods and black-box models. The results are shown as the mean ± std of the five-fold cross-validation experiment. "N/A" indicates invalid data.

Dataset	Model	Disease Diagnosis			Concept Detection	
		AUC (%)	ACC (%)	F1-score (%)	AUC (%)	ACC (%)
Derm7pt	PCBM	81.32 ± 2.12	75.82 ± 2.00	71.10 ± 1.74	N/A	N/A
	PCBM-H	85.87 ± 1.53	78.60 ± 2.79	75.50 ± 3.09	N/A	N/A
	CBIVLM	83.45 ± 3.59	84.13 ± 2.71	71.24 ± 2.54	N/A	N/A
	ECBM	75.03 ± 2.06	76.43 ± 2.47	73.50 ± 2.24	70.59 ± 2.60	78.85 ± 1.40
	AEC	91.27 ± 2.02	84.88 ± 2.05	80.99 ± 3.32	76.61 ± 1.61	75.30 ± 0.72
	CBM	92.88 ± 1.90	85.89 ± 1.92	82.18 ± 2.67	82.15 ± 2.68	80.00 ± 1.87
	CCBM	**93.15 ± 3.68**	**87.18 ± 1.18**	**84.20 ± 4.90**	**83.07 ± 1.53**	**80.71 ± 0.87**
	Inception-v3	92.02 ± 2.53	86.46 ± 2.54	83.13 ± 3.31	N/A	N/A
Skincon	PCBM	69.06 ± 1.23	72.48 ± 1.56	39.55 ± 0.97	N/A	N/A
	PCBM-H	72.85 ± 1.66	68.42 ± 3.07	53.33 ± 3.38	N/A	N/A
	CBIVLM	71.34 ± 2.16	70.74 ± 1.11	66.15 ± 1.50	N/A	N/A
	ECBM	68.04 ± 1.68	79.16 ± 1.06	61.39 ± 2.29	64.64 ± 0.97	90.94 ± 0.20
	AEC	83.86 ± 0.61	**80.24 ± 0.52**	63.89 ± 1.63	58.64 ± 0.90	90.64 ± 0.10
	CBM	80.01 ± 1.25	78.42 ± 1.31	60.57 ± 3.04	62.14 ± 1.37	89.32 ± 0.14
	CCBM	**85.15 ± 1.45**	80.22 ± 0.90	**67.08 ± 0.62**	**80.51 ± 0.51**	**91.32 ± 0.14**
	Inception-v3	79.92 ± 1.48	77.52 ± 1.47	59.86 ± 2.78	N/A	N/A
BrEaST	PCBM	75.41 ± 5.74	68.43 ± 7.64	64.79 ± 8.28	N/A	N/A
	PCBM-H	79.63 ± 2.95	70.02 ± 3.87	67.73 ± 3.81	N/A	N/A
	ECBM	68.09 ± 7.50	68.83 ± 6.10	66.76 ± 6.24	59.20 ± 1.93	77.94 ± 1.65
	AEC	82.80 ± 3.32	72.40 ± 3.39	68.60 ± 3.51	69.71 ± 1.52	77.82 ± 1.70
	CBM	87.42 ± 4.27	77.21 ± 8.62	76.29 ± 8.31	70.76 ± 1.38	77.49 ± 1.43
	CCBM	**88.89 ± 3.67**	**78.81 ± 6.60**	**76.73 ± 7.12**	**74.31 ± 3.32**	**81.03 ± 2.16**
	ResNet50	86.97 ± 6.14	77.61 ± 6.23	76.39 ± 6.09	N/A	N/A

3.3.2. Ablation Study

To further explore the impact of the discovered local concepts on model performance, we conducted ablation experiments on the *Derm7pt* and *BrEaST* datasets. The results are summarized in Table 3. The CCBM's performance remains competitive with other explainable methods even when the local concept learning branch is removed, especially for the concept detection tasks. Moreover, with the local concept learning, the performance of CCBM on the two tasks is significantly improved, which means the discovered local concepts revise the weight distribution of predefined concepts and provide more information to improve model predictions in conjunction with the existing predefined concepts. Specifically, for the *Derm7pt* and *BrEaST*, which have limited predefined global concepts, there is potential to find more valuable concepts from the related medical domain. Hence, more significant improvements can be observed in these two datasets' classification and concept detection tasks. Even for the dataset with enough predefined concepts, the *Skincon*, CCBM can still further improve the accuracy for image classification and concept detection.

Table 3: Results of CCBM without (✗) and with (✓) local concept learning branch. The results are shown as the mean ± std of the five-fold cross-validation experiments.

Dataset	Local Concept Branch	Disease Diagnosis			Concept Detection	
		AUC (%)	ACC (%)	F1-score (%)	AUC (%)	ACC (%)
Derm7pt	✗	91.68 ± 1.76	83.92 ± 2.91	81.52 ± 3.22	82.78 ± 2.28	80.30 ± 1.78
	✓	**93.15 ± 3.68**	**87.18 ± 1.18**	**84.20 ± 4.90**	**83.07 ± 1.53**	**80.71 ± 0.87**
Skincon	✗	84.41 ± 1.20	**80.25 ± 1.20**	**67.34 ± 2.72**	79.27 ± 0.75	91.31 ± 0.20
	✓	**85.15 ± 1.45**	80.22 ± 0.90	67.08 ± 0.62	**80.51 ± 0.51**	**91.32 ± 0.14**
BrEaST	✗	88.16 ± 4.52	73.57 ± 14.07	72.79 ± 14.17	66.49 ± 4.72	79.52 ± 4.26
	✓	**88.89 ± 3.67**	**78.81 ± 6.60**	**76.73 ± 7.12**	**74.31 ± 3.32**	**81.03 ± 2.16**

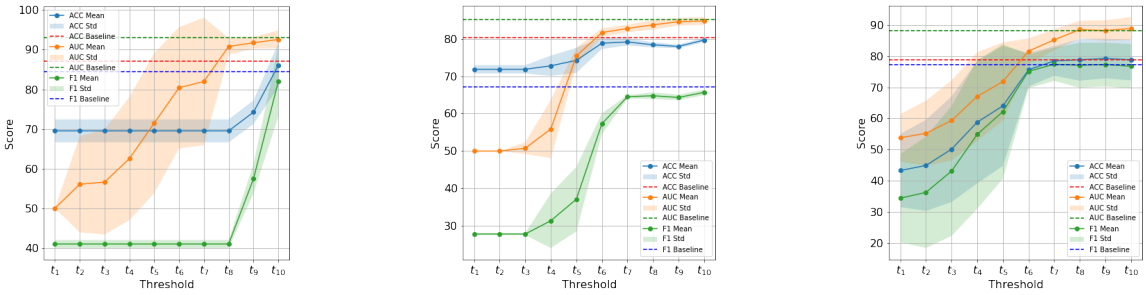
3.4. Explanability Analysis

3.4.1. Inference-time Intervention for Faithfulness

Faithfulness indicates that the model explanations could faithfully elucidate the model decision. In the inference-time intervention experiments, we artificially modify partial concept predictions to observe the resulting changes in the final model decisions, allowing us to assess the effectiveness of the concept explanations. Specifically, we establish a set of thresholds for concept scores during model inference, where any concept scores surpassing the threshold are reset to zero. Subsequently, we present the diagnosis outcomes inferred using the intervened concept scores. Fig. 2 shows that the disease diagnosis performance is notably impacted by the intervention of concept scores in five-fold cross-validation experiments. For these three datasets, the AUC, ACC, and F1-score generally exhibit significant improvements as the intervention threshold increases. When the threshold is set too low, the model performance deteriorates (with AUC dropping to 50%). These intervention experiments underscore the model's heavy reliance on predicted concept scores and affirm the faithfulness of the explanations provided. For the datasets with a small size, the effect of interference on the data results of different folds will be large, so there is a larger variance observed in the results from the *Derm7pt* and *BrEaST* datasets.

3.4.2. Visual and Textual Explanations for Plausibility

How to explain model decisions is a critical issue in explainable artificial intelligence, encompassing the necessity for model explanations to be both convincing and understandable to humans in real-world scenarios. As shown in Fig. 3, we showcase the explanations for one image from the *Derm7pt* and give the statistical information of the discovered concepts on the *Derm7pt* dataset. It can be observed that, for visual explanations, CCBM can capture the important areas for recognizing the specific predefined concepts and generate the corresponding concept activation maps using Grad-CAM (Selvaraju et al., 2020) and give the concept scores. Additionally, we can generate textual summaries via CCBM to report the concepts in the image and the top-k discovered local concepts. These can provide a decision-making basis for the model and improve its reliability. Moreover, we visualize the distribution of the discovered top-1 concepts of training and testing samples. It can be



(a) *Derm7pt* (b) *Skincon* (c) *BrEaST*

Figure 2: Inference-time intervention results on three datasets. The x-axis represents the thresholds ($t_1 \leq t_2 \leq ... \leq t_{10}$), and the y-axis represents the diagnosis performance after intervention. (a) Inference-time intervention results on Derm7pt. (b) Inference-time intervention results on BrEaST. (c) Inference-time intervention results on Skincon.

found that the first four concepts of the training data and the test data are the same, although their rankings are not completely consistent. These concepts are highly related to the characteristics of melanoma. The results reflect that the concepts we discovered can indicate key features to distinguish the categories and guide model predictions.

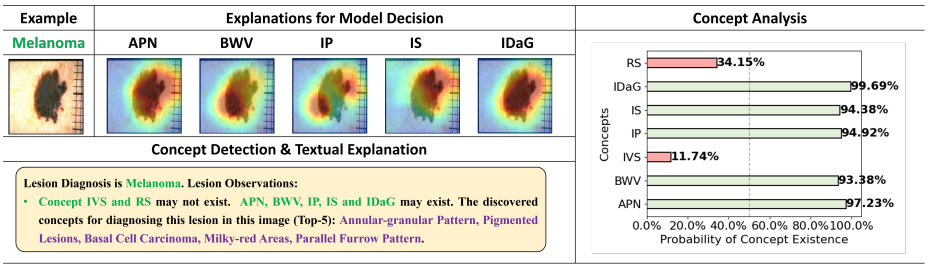


Figure 3: Visual and textual explanations of one image in Derm7pt, all of whose labels are predicted correctly, and the discovered concept analysis for the *Derm7pt* dataset, where the bar images show the distributions of the top-1 discovered concepts.

4. Conclusion

In this paper, we introduce a concept complement bottleneck model for interpretable medical image diagnosis. CCBM discovers important concepts from a concept bank while simultaneously learning predefined concepts for disease diagnosis. By incorporating concept adapters with visual-text concept cross-attention modules, we create a more robust concept bottleneck model that enhances the precision and effectiveness of disease predictions. An effective strategy is proposed for learning unknown local concepts, which aims to extract more significant information to improve model performance. Comprehensive experiments demonstrate that our model achieves superior classification performance in both concept detection and disease diagnosis tasks, providing more faithful and interpretable explanations.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5948–5955, 2023.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao. Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 700–708, 2020.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1820–1828, 2021.

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

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 5338–5348. PMLR, 2020.

Weiji Kong, Xun Gong, and Juan Wang. Lce: A framework for explainability of dnns for ultrasound image based on concept discovery. *arXiv preprint arXiv:2408.09899*, 2024.

Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2020.

Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ugne Klimiene, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, Markus Denzinger, David Niederberger, Sven Wellmann,

Ece Ozkan, et al. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, 2024.

Mert Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Cristiano Patrício, Luís F Teixeira, and João C Neves. Towards concept-based interpretability of skin lesion diagnosis using vision-language models. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023.

Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1): 148, 2024.

Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10295, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2020.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11040, 2024.

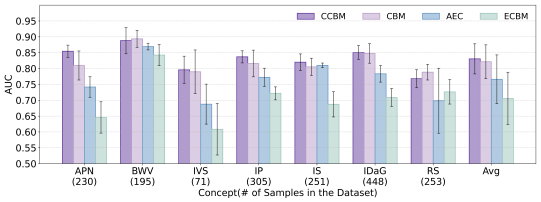
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *The Twelfth International Conference on Learning Representations*, 2024.

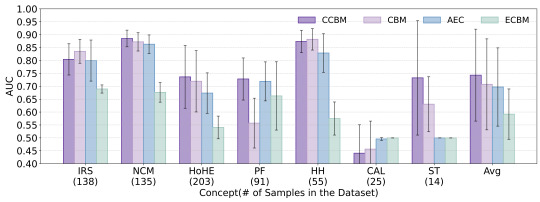
Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages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

Appendix A. Details of concept detection

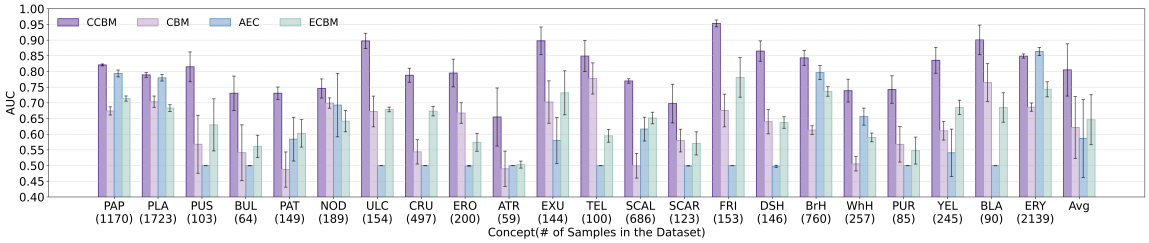
Fig. 4 displays the fine-grained evaluation results of CCBM and other competitors on the concept detection task. CCBM achieves the highest average AUCs for the three datasets.



(a) Concept AUC on Derm7pt



(b) Concept AUC on BrEaST



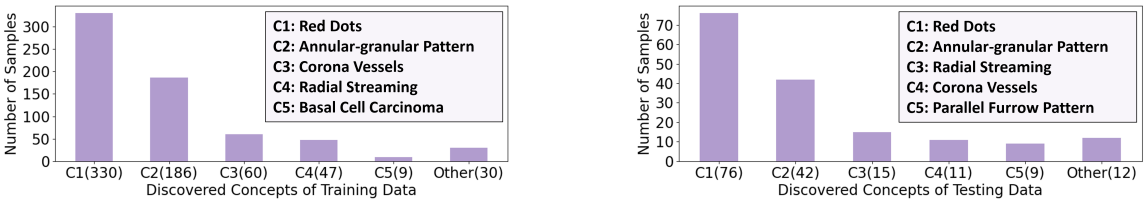
(c) Concept AUC on Skincon

Figure 4: The fine-grained results of the concept detection task on the *Derm7pt*, *BrEaST*, and *Skincon* datasets. The results are the means and stds of the five-fold cross-validation experiments. The "Avg" is the mean and std of the concept AUCs over all the concepts.

Particularly, our CCBM demonstrates strong performance across most concepts, while CBM and other models exhibit significant misclassification for some concepts with fewer positive samples. For example, for the *Skincon*, the results of "BUL", "PUR", "ATR", and "BLA" are representative. While the number of their positive samples is smaller than one hundred, the concept accuracy of CCBM is higher by 10% to 20% than the sub-optimal models. It indicates that CCBM learns each concept more effectively and fairly, benefiting from the concept adapters and independent bottleneck layers.

Appendix B. The visualization for the discovered concepts

We visualized the distribution of the discovered top-1 concepts of training and testing samples as shown in Fig. 5. It can be found that the first four concepts of the training data and the test data are the same, although their rankings are not completely consistent. These concepts are highly related to the characteristics of melanoma, which reflects that the concepts we discovered can indicate key features to distinguish the categories and guide model predictions. With the local branch, the weights of the concepts ("Irregular Vascular Structures (IVS)" and "Irregular Dots and Globules") decreased significantly. The discovered concepts, "Red Dots" and "Corona Vessel", are highly related to the above concepts semantically, but are not the same. This means that CCBM discovers text concepts that are more beneficial for the classification tasks. Additionally, "Annular-granular Pattern" and "Radial Streaming" are also highly specific dermoscopic signs of melanoma. Specifically, the concept "Annular-granular Pattern" is highly indicative of lentigo maligna, which is a type of melanoma in situ. The "Radial Streaming" indicates focal irregular streaking or peripheral linear projections.



(a) Discovered Concepts from Training

(b) Discovered Concepts from Testing

Figure 5: The discovered concept analysis for the *Derm7pt* dataset. (a) The statistical results of the top-1 discovered concepts of the training samples on the *Derm7pt*. (b) The statistical results of the top-1 discovered concepts of the testing samples on the *Derm7pt*.