SANT-CBM: Structurally-Aware and Noise-Tolerant Semi-supervised Concept Bottleneck Models

Hongwei Liu[1], Jia Liu[1], and Songning Lai[2] *

[1] Shanghai normal university
{1000552907,1000554321}@smail.shnu.edu.cn
[2] HKUST(GZ)
lais0328eee@gmail.com

Abstract. Concept Bottleneck Models (CBMs) have emerged as a promising direction for interpretable medical image analysis by grounding predictions on human-understandable concepts. However, training effective CBMs typically requires dense concept annotations by clinical experts, which is prohibitively expensive and scarce. While semi-supervised CBMs attempt to leverage unlabeled data, they often suffer from **label noise** inherent in pseudo-labeling and fail to preserve the **pathological correlations** between concepts (e.g., symptom co-occurrences). In this paper, we propose a novel **Structurally-Aware and Noise-Tolerant CBM (SANT-CBM)** framework for semi-supervised medical image classification. To address the unreliability of pseudo-labels, we introduce a **GMM-based Dynamic Weighting** mechanism that models the loss distribution of unlabeled samples to separate clean supervision from noise, effectively "denoising" the training process. Furthermore, we propose a **Structural Consistency Loss** that aligns the correlation matrix of predicted concepts on unlabeled data with the reference correlation derived from clinical priors. This ensures that the model learns biologically plausible concept combinations rather than independent features. Extensive experiments on two public dermatology datasets (Fitzpatrick17k and 7-point) demonstrate that our method significantly outperforms state-of-the-art semi-supervised CBM baselines. Specifically, our approach achieves diagnostic accuracies of **77.83%** (vs. 71.52%) on Fitzpatrick17k and **73.27%** (vs. 67.52%) on 7-point (based on 10% labeled), establishing a new benchmark for label-efficient and interpretable medical diagnosis.

Keywords: Concept Bottleneck Models · Semi-supervised Learning · Medical Image Classification.

1 Introduction

Deep learning has revolutionized medical image analysis [9,21], yet the opaque "black-box" nature of these models hinders their adoption in high-stakes clinical

* Correspondence to Songning Lai {lais0328eee@gmail.com}.

decision-making [15]. To engender trust, Concept Bottleneck Models (CBMs) [10] have emerged as a dominant interpretable paradigm. By decomposing prediction into a two-stage process—mapping images to human-understandable concepts (e.g., "irregular streaks") before making a diagnosis—CBMs emulate clinical reasoning and enable expert intervention [7,8].

However, the efficacy of CBMs relies heavily on dense, expert-annotated concept labels, which are prohibitively expensive to acquire compared to global diagnostic labels. In dermatology, for instance, delineating fine-grained dermoscopic features for thousands of lesions is impractical, resulting in datasets like Fitzpatrick17k [5] having abundant images but scarce concept annotations. To address this, Semi-supervised Learning (SSL) has been introduced to CBMs. Notably, the recent state-of-the-art SSCBM [6] utilizes unlabeled data by generating pseudo-labels via K-Nearest Neighbors (KNN) [14] and enforcing visual alignment through saliency maps.

While SSCBM represents a significant step forward, directly applying such frameworks to medical diagnosis reveals two critical limitations regarding reliability and plausibility: **First, vulnerability to label noise.** Existing methods like SSCBM typically rely on heuristics (e.g., KNN or fixed thresholds) to generate pseudo-labels. These deterministic approaches lack a probabilistic mechanism to filter noise. In the early training stages, incorrect pseudo-labels can propagate errors, leading to *confirmation bias* where the model overfits to its own hallucinations rather than learning true features. **Second, neglect of pathological structure.** SSCBM focuses on *local* alignment between image pixels and individual concepts (visual grounding) but overlooks the *global* correlations between concepts. In medical reality, symptoms are not independent variables; they form specific syndromes (e.g., pigment networks "strongly co-occur with specific globules"). Ignoring these intrinsic pathological priors can lead to biologically implausible predictions—where a model might predict a combination of concepts that never co-exist in clinical practice.

To bridge these gaps, we propose **SANT-CBM**, a Structurally-Aware and Noise-Tolerant framework. Unlike previous works that focus solely on visual alignment, we argue that robust medical SSL requires adherence to both data cleanliness and clinical logic. Our contributions are three-fold:

1. We introduce a **GMM-based Dynamic Weighting** mechanism. Instead of relying on hard thresholds or KNN, we model the loss distribution of unlabeled samples using a Gaussian Mixture Model to probabilistically distinguish "clean" supervision from noise, effectively denoising the training process.
2. We propose a novel **Structural Consistency Loss** that integrates clinical priors. By aligning the correlation matrix of predicted concepts with ground-truth pathological dependencies, we constrain the model to learn biologically plausible concept combinations rather than independent features.
3. Extensive experiments on Fitzpatrick17k and 7-point datasets demonstrate that SANT-CBM significantly outperforms the state-of-the-art SSCBM. Specifically, we achieve diagnostic accuracy improvements of **6.31%** and **5.75%**

respectively (based on 10% labeled), establishing a new benchmark for label-efficient medical diagnosis.

2 Related Work

2.1 Concept Bottleneck Models

Concept Bottleneck Models (CBMs) [10] have emerged as a dominant paradigm for interpretable medical diagnosis by explicitly mapping raw images to high-level clinical concepts before making final predictions. To improve the expressiveness of simple binary concepts, recent advancements like Concept Embedding Models (CEM) [3] utilize high-dimensional concept embeddings, achieving performance comparable to black-box models. Despite these architectural improvements, the scalability of CBMs is severely constrained by the *annotation bottleneck*. Training robust CBMs typically necessitates dense, expert-verified concept labels. While some works explore unsupervised or LLM-guided concept discovery [13,20], these methods often yield concepts that lack the rigorous clinical definition required for medical standards. Consequently, developing label-efficient CBMs that can leverage sparse annotations without compromising diagnostic precision remains an open challenge.

2.2 Semi-supervised Learning in Medical Imaging

Semi-supervised Learning (SSL) has been extensively applied in medical imaging to mitigate data scarcity, with consistency regularization (e.g., FixMatch [16]) and pseudo-labeling being standard approaches. Recently, **SSCBM** [6] extended SSL to the CBM framework. It generates pseudo-labels via KNN and enforces *visual alignment* by constraining concept saliency maps to match image features. However, two critical limitations persist in current state-of-the-art methods: **(1) Susceptibility to Noise:** Heuristic pseudo-labeling methods (like KNN used in SSCBM) are deterministic and prone to *confirmation bias*, especially when the initial labeled set is small or unrepresentative. They lack a probabilistic mechanism to filter out noisy pseudo-labels during training. **(2) Neglect of Structural Logic:** SSCBM focuses on *local* alignment (pixel-to-concept) but overlooks *global* pathological correlations (concept-to-concept). In medical diagnosis, clinical findings are structurally correlated (e.g., specific pigment networks co-occur with globules). Ignoring these priors leads to biologically implausible explanations. Our SANT-CBM addresses these gaps by integrating a GMM-based noise tolerance module and a structural consistency loss, ensuring that the model learns both visually grounded and pathologically logical concepts.

3 Method

3.1 Preliminaries

We address the problem of semi-supervised medical image classification with interpretable concepts. Let $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{c}_i)\}_{i=1}^{N_L}$ denote the labeled set, where

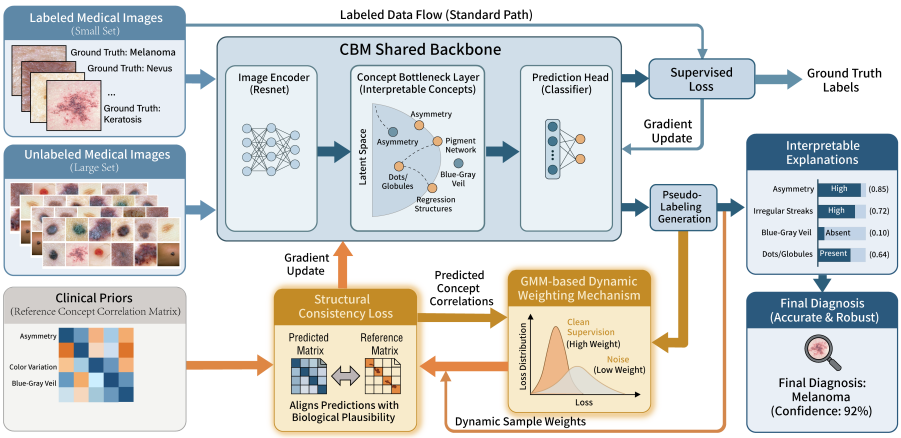

Fig. 1. Overview of the proposed SANT-CBM framework. It consists of two key branches utilizing unlabeled data: (a) A **Noise-Tolerant** branch that uses GMM to dynamically weight pseudo-labels based on their loss distribution, filtering out noise. (b) A **Structurally-Aware** branch that aligns the concept correlation matrix of unlabeled predictions with a clinical prior reference matrix derived from labeled data.

$\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}$ is the input image, $\mathbf{y}_i \in \{0,1\}^K$ is the diagnostic label, and $\mathbf{c}_i \in \{0,1\}^M$ is the vector of M binary clinical concepts. Additionally, a large unlabeled set $\mathcal{D}_U = \{(\mathbf{u}_i)\}_{i=1}^{N_U}$ is provided, where $N_U \gg N_L$.

Building upon the CBM framework, our model consists of two sequential mappings: a concept predictor $g(\cdot) : \mathbf{x} \to \hat{\mathbf{c}}$ and a label predictor $h(\cdot) : \hat{\mathbf{c}} \to \hat{\mathbf{y}}$. For the labeled set $\mathcal{D}_L$, the model is optimized via a supervised loss $\mathcal{L}_{sup}$, which combines cross-entropy (CE) for diagnosis and binary cross-entropy (BCE) [19] for concept prediction:

$$\mathcal{L}_{sup} = \frac{1}{N_L}\sum_{i=1}^{N_L}[\mathcal{L}_{CE}(h(g(\mathbf{x}_i)),\mathbf{y}_i) + \lambda_c\mathcal{L}_{BCE}(g(\mathbf{x}_i),\mathbf{c}_i)] \tag{1}$$

where λ_c balances the task and concept supervision.

3.2 GMM-based Dynamic Weighting for Denoising

The core challenge in semi-supervised CBMs is the quality of pseudo-labels for $\mathcal{D}_U$. Naive approaches (e.g., thresholding) are deterministic and often propagate confirmation bias. To address this, we propose a probabilistic noise-tolerant mechanism.

Pseudo-label Generation via Feature Bank. Following recent advances, we maintain a feature bank $\mathcal{B}$ storing features $\mathbf{f}_k$ and concept labels $\mathbf{c}_k$ from $\mathcal{D}_L$. For an unlabeled sample $\mathbf{u}_i$, we retrieve its K-nearest neighbors from $\mathcal{B}$ based on

cosine similarity. The initial pseudo-label $\tilde{\mathbf{c}}_i$ is generated by aggregating neighbor concepts using Softmax-normalized similarity weights:

$$\tilde{\mathbf{c}}_i = \sum_{k \in \mathcal{N}_K(\mathbf{u}_i)} \frac{\exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_k)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_j)/\tau)} \cdot \mathbf{c}_k \tag{2}$$

where τ is a temperature parameter.

Probabilistic Noise Filtering. Instead of directly trusting $\tilde{\mathbf{c}}_i$, we observe that during training, the model tends to learn clean patterns faster than noisy ones ("small-loss trick"). We model the alignment loss distribution for each concept j independently. Let $\ell_{i,j} = (g(\mathbf{u}_i)_j - \tilde{c}_{i,j})^2$ be the squared error for concept j of sample i. We fit a two-component Gaussian Mixture Model (GMM) [12] to the distribution of losses $\ell_{i,j}$ over the unlabeled set:

$$p(\ell|j) = \pi_j \mathcal{N}(\ell|\mu_{j,\text{clean}}, \sigma^2_{j,\text{clean}}) + (1-\pi_j)\mathcal{N}(\ell|\mu_{j,\text{noisy}}, \sigma^2_{j,\text{noisy}}) \tag{3}$$

where $\mu_{j,\text{clean}} < \mu_{j,\text{noisy}}$. The "cleanliness" weight $w_{i,j}$ is then derived as the posterior probability that the loss belongs to the clean component:

$$w_{i,j} = p(\text{clean}|\ell_{i,j}) = \frac{\pi_j \mathcal{N}(\ell_{i,j}|\mu_{j,\text{clean}}, \sigma^2_{j,\text{clean}})}{p(\ell_{i,j}|j)} \tag{4}$$

This allows the model to dynamically down-weight unreliable pseudo-labels for specific concepts while retaining useful signals. The denoised alignment loss is:

$$\mathcal{L}_{denoise} = \frac{1}{N_U} \sum_{i=1}^{N_U} \sum_{j=1}^{M} w_{i,j} \cdot (g(\mathbf{u}_i)_j - \tilde{c}_{i,j})^2 \tag{5}$$

3.3 Structural Consistency with Clinical Priors

Visual alignment alone ensures local correctness but ignores global pathological [2] logic (e.g., symptom A and B should strictly co-occur). We enforce structural consistency by aligning the correlation matrix of predictions with a clinical prior.

Reference Matrix Construction. Prior to training, we compute a static Reference Correlation Matrix $\mathbf{R}_{ref} \in \mathbb{R}^{M \times M}$ using the ground-truth concepts from $\mathcal{D}_L$. $\mathbf{R}_{ref}[m, n]$ represents the Pearson correlation [18] coefficient between concept m and n, encapsulating the "clinical common sense."

Differentiable Batch Correlation Alignment. During training, for a batch of unlabeled images, we compute the predicted concept matrix $\mathbf{P} \in \mathbb{R}^{B \times M}$. To ensure differentiability, we calculate the batch-wise Pearson correlation matrix

$\hat{\mathbf{R}}_{batch}$ using native matrix operations. For any two predicted concept vectors $\mathbf{p}_m, \mathbf{p}_n \in \mathbb{R}^B$:

$$\hat{\mathbf{R}}_{batch}[m,n] = \frac{\sum_{b=1}^{B}(p_{b,m} - \bar{p}_m)(p_{b,n} - \bar{p}_n)}{\sqrt{\sum_{b=1}^{B}(p_{b,m} - \bar{p}_m)^2}\sqrt{\sum_{b=1}^{B}(p_{b,n} - \bar{p}_n)^2} + \epsilon} \tag{6}$$

where $\bar{p}$ denotes the mean within the batch and ϵ is a small constant for stability. Finally, we impose the Structural Consistency Loss using the Frobenius norm [1] to align the predicted structure with the clinical prior:

$$\mathcal{L}_{struct} = \left\| \mathbf{R}_{ref} - \hat{\mathbf{R}}_{batch} \right\|_F^2 = \sum_{m=1}^{M}\sum_{n=1}^{M}\left(\mathbf{R}_{ref}[m,n] - \hat{\mathbf{R}}_{batch}[m,n]\right)^2 \tag{7}$$

This constraint penalizes biologically implausible predictions (e.g., predicting two mutually exclusive symptoms simultaneously), effectively regularizing the model with domain knowledge.

To formulate the overall objective, we combine the supervised loss, the noise-tolerant pseudo-labeling loss, and the structural consistency loss as $\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u \mathcal{L}_{denoise} + \lambda_s \mathcal{L}_{struct}$, where λ_u and λ_s are hyperparameters balancing the semi-supervised and structural components. Additionally, we employ a warm-up strategy for the GMM weighting to ensure training stability in early epochs.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our framework on two public dermatology datasets with sparse concept annotations:

- **Fitzpatrick17k** [5]: A large-scale dataset containing 16,577 clinical images across 114 skin conditions. Following standard protocols, we group them into benign, malignant, Non-neoplastic categories. We select 22 concepts (*i.e.*, K = 22) from the F17k dataset that are represented by at least 50 images for our analysis (e.g., texture, color).
- **7-point Checklist** [17]: A dermoscopy dataset comprising 1,011 images. It provides diagnosis labels (e.g., Melanoma, Seborrheic Keratosis) and dense annotations for 19 dermoscopic concepts utilized in the 7-point scoring system.

For semi-supervised settings, we partition the training set into labeled ($\mathcal{D}_L$) and unlabeled ($\mathcal{D}_U$) subsets with varying ratios (e.g., 10%, 20%), while keeping the validation and test sets fixed.

Baselines. We compare SANT-CBM against these methods: (1) **Supervised CBM**: Standard **CBM** [10] and **CEM** [3]. (2) **Semi-supervised CBMs**: The state-of-the-art **SSCBM** [6], which represents the most direct competitor.

Table 1. Comparison with state-of-the-art methods on Fitzpatrick17k and 7-point datasets. We report Task Accuracy (%) and Concept Accuracy (AUC %). Best results in bold.

Method	Backbone	Fitzpatrick17k				7-point Checklist			
		10% Labeled		20% Labeled		10% Labeled		20% Labeled	
		Concept	Task	Concept	Task	Concept	Task	Concept	Task
Standard CBM [10]	ResNet-50	86.31	71.51	87.27	72.53	63.56	69.43	65.31	70.67
CEM [3]	ResNet-50	86.26	72.31	87.57	73.19	64.71	68.15	66.37	69.23
SSCBM [6]	ResNet-50	88.54	71.52	88.79	72.03	**71.95**	67.52	**72.57**	67.97
SANT-CBM (Ours)	ResNet-50	**88.84**	**77.83**	**88.97**	**78.26**	71.65	**73.27**	72.31	**72.97**

Implementation Details. We employ **ResNet-50** pretrained on ImageNet as the feature extractor. The concept predictor is a 2-layer MLP with ReLU activation. We use the **sgd** [11] optimizer with a learning rate of 5e-3 and a batch size of 16. The hyperparameters in Eq. (8) are set to $\lambda_u = 1.1$ and $\lambda_s = 0.3$. The GMM warm-up period is set to 6 epochs on average. All experiments are implemented in PyTorch on NVIDIA RTX 3090 GPUs. Results are reported as the mean of 3 independent runs.

4.2 Comparison with State-of-the-Arts

Table 1 presents the performance comparison on Fitzpatrick17k and 7-point datasets under different labeled data ratios. Our SANT-CBM essentially outperforms all semi-supervised baselines. Notably, in the low-data regime (10% labels), SANT-CBM surpasses the SOTA SSCBM by 0.3% in Concept Accuracy and 6.3% in Task Accuracy on Fitzpatrick17k. While SSCBM improves visual feature alignment, its performance plateaus due to noisy pseudo-labels. In contrast, our method effectively filters out unreliable supervision via GMM, approaching the performance of fully supervised models using only a fraction of annotations.

4.3 Ablation Studies

To validate the contribution of each component, we conduct ablation studies on the Fitzpatrick17k dataset with 10% labels (Table 2). The baseline, relying on naive pseudo-labeling, clearly suffers from noise propagation. **First**, integrating the GMM-based dynamic weighting ($\mathcal{L}_{denoise}$) improves Concept AUC by 1.5%, indicating that down-weighting high-loss samples effectively prevents the model from memorizing incorrect pseudo-labels. **Second**, independently applying the Structural Consistency Loss ($\mathcal{L}_{struct}$) yields a substantial 3.6% gain in Task Accuracy. By enforcing the prediction correlation matrix to align with the clinical prior $\mathbf{R}_{ref}$, the model is constrained to learn pathologically meaningful concept combinations, which intrinsically benefits the final diagnosis. **Finally**, combining both modules (SANT-CBM) yields the highest overall performance (77.83% Task Acc. and 88.84% Concept Acc), demonstrating that probabilistic noise filtering and global structural regularization are highly complementary.

Table 2. Ablation study on Fitzpatrick17k (10% labels).

GMM Denoising	Structural Loss	Concept Acc (%)	Task Acc (%)
-	-	86.31	71.51
✓	-	87.81	74.38
-	✓	86.47	75.13
✓	✓	**88.84**	**77.83**

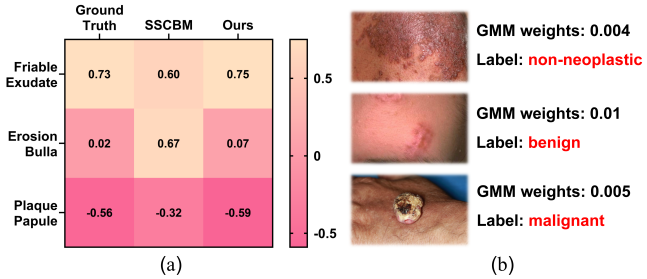

Fig. 2. (a) Visualization of Concept Correlation Matrices: Ground Truth vs. Baseline vs. Ours. (b) Samples identified as "Noisy" by our GMM module (Low Weight).

4.4 Qualitative Analysis and Visualization

To intuitively understand how SANT-CBM improves interpretability, we provide visualizations in Fig. 2. **Concept Correlation Alignment:** Fig. 2(a) compares the concept correlation matrices. The baseline model (SSCBM) predicts a noisy matrix where unrelated concepts show high correlation (e.g., Erosion and Bulla). In contrast, SANT-CBM produces a structure that closely mirrors the Ground Truth reference, confirming that our model learns valid clinical logic. **Noise Identification:** Fig. 2(b) displays samples with low GMM weights ($w_i < 0.1$). We observe that these images are often blurry, contain hair occlusion, or have ambiguous boundaries. By assigning low weights to these reliable samples, GMM effectively protects the model from learning wrong features.

5 Conclusion

In this paper, we proposed **SANT-CBM**, a robust framework for semi-supervised concept bottleneck models that tackles the critical issues of pseudo-label noise and structural inconsistency. By synergizing GMM-based dynamic denoising with a differentiable clinical prior alignment, our approach ensures that concept predictions are both statistically reliable and pathologically plausible. Experiments on two dermatology datasets demonstrate that SANT-CBM significantly outperforms state-of-the-art baselines in label-scarce regimes. Future work will explore integrating Large Language Models (LLMs) [4] to inject external medical knowledge graphs, further reducing dependency on dataset statistics and enhancing generalization.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, C., Qi, J., Liu, X., Bin, K., Fu, R., Hu, X., Zhong, P.: Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26836–26845 (2024)
2. Deng, R., Liu, Q., Cui, C., Yao, T., Xiong, J., Bao, S., Li, H., Yin, M., Wang, Y., Zhao, S., et al.: Hats: Hierarchical adaptive taxonomy segmentation for panoramic pathology image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 155–166. Springer (2024)
3. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in neural information processing systems **35**, 21400–21413 (2022)
4. Fang, X., Lin, Y., Zhang, D., Cheng, K.T., Chen, H.: Aligning medical images with general knowledge from large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 57–67. Springer (2024)
5. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: 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. pp. 1820–1828 (2021)
6. Hu, L., Huang, T., Xie, H., Gong, X., Ren, C., Hu, Z., Yu, L., Ma, P., Wang, D.: Semi-supervised concept bottleneck models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2110–2119 (2025)
7. Hu, L., Lai, S., Chen, W., Xiao, H., Lin, H., Yu, L., Zhang, J., Wang, D.: Towards multi-dimensional explanation alignment for medical classification. Advances in Neural Information Processing Systems **37**, 129640–129671 (2024)
8. Hu, L., Lai, S., Hua, Y., Yang, S., Zhang, J., Wang, D.: Stable vision concept transformers for medical diagnosis. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 317–332. Springer (2025)
9. Jiang, X., Hu, Z., Wang, S., Zhang, Y.: Deep learning for medical image-based cancer diagnosis. Cancers **15**(14), 3608 (2023)
10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
11. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 552–561. Springer (2020)
12. Li, Z., Yu, F., Lu, J., Qian, Z.: Gmm-coregnet: A multimodal groupwise registration framework based on gaussian mixture model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 629–639. Springer (2024)
13. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023)

14. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
15. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
16. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
17. Walter, F.M., Prevost, A.T., Vasconcelos, J., Hall, P.N., Burrows, N.P., Morris, H.C., Kinmonth, A.L., Emery, J.D.: Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study. The British Journal of General Practice **63**(610), e345 (2013)
18. Woodland, M., Castelo, A., Al Taie, M., Albuquerque Marques Silva, J., Eltaher, M., Mohn, F., Shieh, A., Kundu, S., Yung, J.P., Patel, A.B., et al.: Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 87–97. Springer (2024)
19. Wu, W., Qiu, X., Song, S., Chen, Z., Huang, X., Ma, F., Xiao, J.: Prompt categories cluster for weakly supervised semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3198–3207 (2025)
20. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023)
21. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)