MDPNet: A Multi-Scale Deformable Prototype Network for Accurate and Interpretable Medical Image Classification

Qinglan Yan[1], Fei Gu[1*], Wenlin Li[2], Chaozhuo Li[3], Shikun Sun[4], Jing Wu[1], and Changsheng Ma[4]

[1] Soochow University, Suzhou, China
[2] Suzhou University of Technology, Suzhou, China
[3] Beijing University of Posts and Telecommunications, Beijing, China
[4] The First Affiliated Hospital of Soochow University, Suzhou, China
gufei@suda.edu.cn

Abstract. There is an increasing demand for medical image classifiers that deliver both high diagnostic accuracy and clinically actionable interpretability. Existing post-hoc explanation methods, such as saliency maps and Grad-CAM, can highlight discriminative regions but often provide limited insight into the underlying decision logic, which can reduce clinical confidence. Prototype learning offers a promising case-based alternative; however, conventional prototype-based models are often limited by single-scale feature representations and sensitivity to class imbalance, which can degrade accuracy and yield less stable explanations. To address these limitations, we propose MDPNet, a Multi-Scale Deformable Prototype Network that improves classification performance while providing prototype-grounded interpretability. MDPNet integrates a multi-scale feature extraction module to capture both fine-grained morphological patterns and global contextual semantics, and introduces a deformable prototype learning mechanism that adaptively aligns prototypes with spatial variability in clinical images. In addition, we adopt a logit-adjusted loss to mitigate the impact of class imbalance during training. Extensive experiments show that MDPNet achieves 93.28±1.54% accuracy on a private premature ventricular contraction (PVC) dataset, 97.90±0.37% accuracy on the public OCT dataset, and 87.65±2.54% accuracy on the public CT dataset, consistently outperforming strong CNN/Transformer baselines and recent prototype-learning approaches.

Keywords: Interpretability · Medical Image Classification · Prototype Learning · Premature Ventricular Contractions · Optical Coherence Tomography.

1 Introduction

Deep learning frameworks have demonstrated impressive performance in learning complex patterns from medical data. A typical model (e.g., a convolutional

neural network) is optimized end-to-end to map an input image to a diagnostic label, and these approaches have achieved strong results across modalities, such as ECG images, MRI, and OCT analysis [1, 2, 5, 16]. Despite these successes, learned representations typically function as opaque black boxes, where decision logic is embedded in complex nonlinear transformations and is difficult to audit. Consequently, unlike evidence-based medical practice, standard predictions produced by these models frequently lack transparency, case-aligned rationales and are difficult to reconcile with clinical evidence, which limits their suitability for high-stakes clinical use requiring physician trust [17].

Recent studies suggest that this limitation can be partially mitigated by incorporating interpretable inductive biases. While post-hoc methods such as Grad-CAM [10, 13, 18] are widely used, they may not reflect causal decision factors and can be sensitive to model perturbations [9]. In contrast, prototype learning (comparing an input with a set of learned representative cases) models can provide case-based explanations that are easier to align with clinical reasoning [3, 8, 20, 21, 24]. Despite the promise of prototype learning, many existing prototype-based methods are limited by their spatial rigidity and single-scale representation. Typically, a prototype corresponds to a fixed, contiguous patch extracted at a single feature scale. This design has two key limitations. First, a rigid contiguous patch assumes diagnostic evidence is spatially adjacent, which can miss multi-focal evidence distributed across the image. For instance, ECG diagnosis often relies on spatially separated P-waves and T-waves, while OCT lesions appear as discrete spots. Second, single-scale features struggle to simultaneously preserve fine-grained detail and global context [6]. As a result, existing models may miss subtle lesions at low resolution or include excessive background due to inflexible receptive fields.

We argue that reliable interpretability in complex medical scenarios requires (i) multi-scale representations that capture both local details and global semantics, and (ii) spatial flexibility that allows different parts of a prototype to deform and align to non-adjacent evidence regions. To this end, we propose MDPNet, an interpretable framework for medical image classification with two key innovations. First, a Multi-Scale Feature Extraction Module (MSFEM) hierarchically captures features, preserving low-level morphological details for tiny anomalies and high-level contextual semantics for global pathological patterns. Second, MDPNet adopts a deformable prototype learning strategy [7] that allows subparts of each prototype to shift independently, enabling a single prototype to represent multiple non-adjacent critical regions. In PVC classification, such evidence may appear across multiple ECG leads and local waveform segments, making flexible evidence matching important for both accuracy and case-level interpretability. This design enables the model to integrate evidence across locations and scales in a unified similarity-based reasoning process, yielding traceable decision rationales that better match diagnostic practice rather than a simple combination of independently acting modules. Moreover, MDPNet incorporates logit adjustment [15] to reduce the impact of class imbalance during training. The key contributions of this work can be summarized as follows:

1. We propose MDPNet, an interpretable medical image classification framework that integrates deformable prototype learning with case-based reasoning.
2. We design a MSFEM module to hierarchically capture both fine-grained morphological details and global contextual patterns, and introduce a deformable prototype learning strategy that adapts prototype matching to spatial variability in clinical images.
3. To mitigate the impact of class imbalance, we incorporate a logit-adjusted loss that calibrates decision boundaries without re-sampling.

2 Methods

2.1 Overview

The overall architecture of the proposed MDPNet is illustrated in Figure 1. The network consists of three key modules: a multi-scale feature extraction module, a deformable prototype learning module, and a classification module.

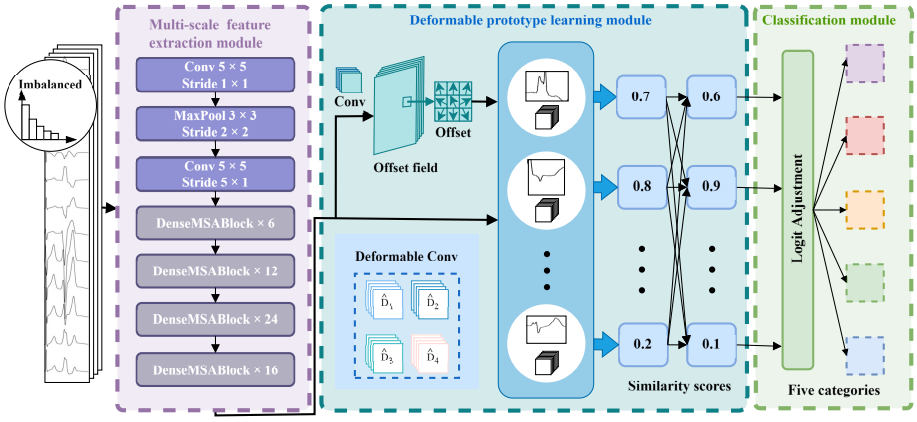

Fig. 1: The illustration of our proposed MDPNet framework for PVC classification in 12-lead ECG images.

2.2 Proposed Architecture

Multi-Scale Feature Extraction Module. Inspired by the multi-scale spatial pyramid attention mechanism [22], we insert a multi-scale feature extraction module into DenseNet bottleneck blocks to enhance multi-scale representations. Formally, given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, we evenly split $\mathbf{F}$

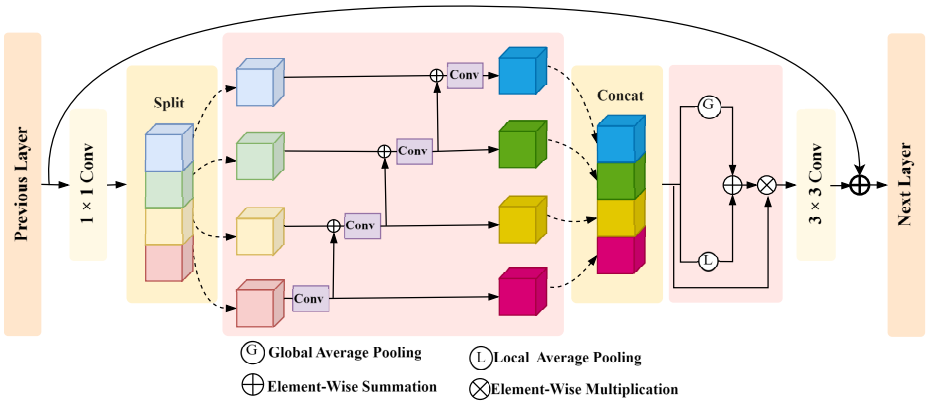

Fig. 2: The illustration of the enhanced DenseNet bottleneck structure integrating a multi-scale feature extraction module for ECG images.

along the channel dimension into n subsets $\mathbf{F}_i \in \mathbb{R}^{c\times H\times W}$, where $c = C/n$. Each subset is processed hierarchically as

$$\mathbf{G}_i = \begin{cases} \text{Conv2D}(\mathbf{F}_i), & i = 1, \\ \text{Conv2D}(\mathbf{F}_i + \mathbf{G}_{i-1}), & 1 < i \leq n. \end{cases} \tag{1}$$

This hierarchical formulation allows each convolution to leverage both current inputs and previously extracted features, progressively enlarging the receptive field and capturing multi-scale ECG patterns.

To adaptively emphasize informative channels, we compute a descriptor for each subset by combining global and local pooling:

$$\mathbf{z}_i = \alpha \cdot \text{GAP}(\mathbf{G}_i) + \beta \cdot \text{LAP}(\mathbf{G}_i), \tag{2}$$

where α and β are learnable weights, $\text{LAP}(\cdot)$ denotes local average pooling implemented as a sliding-window average with kernel size 3×3, stride 1, and padding 1 (i.e., $\text{AvgPool}_{3\times3}$). Each descriptor is passed through a two-layer transformation to produce channel-wise weights that quantify the relative importance of different channels. We normalize these weights across subsets so that multiple scales compete for contribution, preventing redundant amplification across scales. The normalized weights rescale the corresponding feature maps channel-wise and the reweighted features are concatenated along the channel dimension to form the final representation.

Deformable Prototype Learning Module. To accommodate the variability in PVC patterns across 12-lead ECG images, we adopt deformable prototype learning, allowing prototype parts to adapt to different ECG morphologies. A deformable prototype for class c is defined as $\hat{p}^{(c)} \in \mathbb{R}^{\rho_1\times\rho_2\times d}$, where each of the $\rho_1 \times \rho_2$ prototypical parts represents a local semantic pattern in the feature

space. In our setting, we use $\rho_1 = \rho_2 = 5$ and a dilation rate $r = 2$, so that the effective span of an undeformed prototype along each axis is $(\rho_1 - 1)r + 1 = 9$, matching a 9×9 feature map. The offsets are predicted by a convolutional offset head, $\Delta = f_\theta(\hat{z})$, and we regularize them with an L^2 penalty $\lambda\|\Delta\|_2$ to avoid excessive deformation. For out-of-bound sampling, we adopt zero padding, and use bilinear interpolation for fractional sampling locations in the deformable operator. To further ensure interpretability, we periodically perform a push step that updates each prototype by replacing it with the most activated real feature patch from the training set.

During matching, each prototype part computes cosine similarity with the feature map at its offset location, and the responses from all parts are aggregated to form a prototype-level similarity map:

$$k_\theta(\hat{\mathbf{z}})^{(c)}_{a,b} = \sum_{m=1}^{\rho_1} \sum_{n=1}^{\rho_2} \hat{\mathbf{p}}^{(c)}_{m,n} \cdot \hat{\mathbf{z}}_{a+m+\Delta_1,\, b+n+\Delta_2}. \tag{3}$$

Because both prototypes and features are L^2 normalized, the dot product corresponds to cosine similarity. This formulation allows spatial deformation while preserving semantic consistency of the prototype. The final similarity score for each prototype is obtained by max-pooling over all spatial locations:

$$k_\theta(\hat{\mathbf{z}})^{(c)} = \max_{a,b} k_\theta(\hat{\mathbf{z}})^{(c)}_{a,b}. \tag{4}$$

Classification Module. The classification module maps prototype-level similarity scores to class logits. To mitigate class imbalance at the decision level, we apply logit adjustment during training before computing the loss. Specifically, each class logit z_c is adjusted using the empirical class prior, π_c estimated from the training set:

$$z'_c = z_c - \tau \cdot \log \pi_c. \tag{5}$$

We follow Menon et al. [15] and apply logit adjustment during training only, with the temperature parameter τ being tuned on the training set. This adjustment reduces bias toward majority classes when learning decision boundaries. The adjusted logits are used only during training, while the original logits are retained at inference time. We optimize the model with a composite loss:

$$\ell_d(y, \hat{y}) = \ell^{(-)}_{\text{ce}} + \lambda_1 \ell_{\text{clst}} + \lambda_2 \ell_{\text{sep}} + \lambda_3 \ell_{\text{ortho}}, \tag{6}$$

where $\ell^{(-)}_{\text{ce}}$ enlarges inter-class angular margins, and the auxiliary losses encourage alignment with class-specific prototypes, suppress responses to non-matching classes, and promote prototype diversity. We follow the loss-weight settings in [7] unless otherwise specified.

3 Experiments and Results

We evaluated MDPNet on three diverse medical imaging datasets using five-fold cross-validation. The private PVC dataset, collected from The First Affiliated

Hospital of Soochow University under an approved institutional ethics application (Ethics No. 2024320), contains 785 expert-validated, 12-lead ECG images from 486 patients, classified into five anatomical origins: RVOT, LVOT, PM, VA, and HPS. The **OCT dataset (NEH)** [19] consists of 12,649 B-scans across three categories (Normal, Drusen, and CNV) following the preprocessing criteria in [19]. Finally, the **COVID-CT-Dataset** [25] comprises 746 chest CT images formulated as a binary classification task.

3.1 Experimental Setup

We implement MDPNet in PyTorch and train it on a single NVIDIA GeForce RTX 3080Ti GPU with 12GB memory. We use the Adam optimizer with an initial learning rate of 10^{-4}, weight decay of 10^{-3}, batch size 4 for training and batch size 8 for inference. For PVC classification, ECG images are resized to $3 \times 1500 \times 300$, and the multi-scale feature extraction module produces latent representations of size $1024 \times 9 \times 9$. For NEH OCT images, inputs are resized to 512×768 with data augmentation (rotation: $\pm 10°$, translation: up to 15%, scaling: 0.95–1.05, and shearing: $\pm 10°$). For the COVID-CT-Dataset, images are resized and cropped to 224×224 with random resized cropping and horizontal flipping for training, and center cropping for validation. We use 100 prototypes per class for PVC and 50 per class for NEH and CT. We use following metrics to evaluate classification performance: accuracy, precision, recall (sensitivity), and F1-score.

3.2 Overall Performance

Table 1 summarizes MDPNet's classification performance on three datasets. Across ECG, OCT, and CT modalities, MDPNet consistently outperforms or matches strong baselines, including CNN/Transformer models and recent prototype learning approaches. In addition to improved accuracy, MDPNet achieves a more favorable precision-recall balance, indicating better robustness under data heterogeneity and class imbalance, particularly on the multi-class PVC task.

3.3 Ablation Experiments

We conduct ablation experiments(Table 2) to assess the contributions of the Multi-Scale Feature Extraction Module (MSFEM), the Deformable Prototype Learning Module (DPLM), and the Logit Adjustment (LA). On the PVC dataset, MSFEM+DPLM achieves 93.02% accuracy, consistently outperforming MSFEM-only (91.65%) and DPLM-only (91.86%). With all modules enabled, MDPNet reaches 93.28% accuracy with reduced variance across folds (± 1.54), suggesting more stable performance. On the OCT dataset, MSFEM improves accuracy from 96.79% to 97.26%, while DPLM yields a smaller gain (96.79% to 97.18%), likely due to lower spatial variability and a more balanced class distribution. LA provides a larger benefit on PVC than OCT, consistent with the stronger class

Table 1: Overall classification performance (in %) on three datasets.

Model	Accuracy	Recall (Sensitivity)	Precision	F1
PVC (5 classes)				
ResNet50 [11]	76.74±4.04	76.57±8.97	74.03±9.63	74.42±8.66
DenseNet121 [12]	86.37±1.89	86.33±5.62	86.08±4.74	86.05±5.10
CrossViT [4]	83.84±2.90	83.84±9.39	85.25±4.00	84.36±5.50
Swin Transformer [14]	75.13±3.50	50.63±5.01	65.77±6.95	57.01±4.61
Deformable ProtoPNet [7]	91.86±2.06	92.08±4.51	90.59±5.00	91.18±4.40
MedMamba [23]	89.90±1.86	89.54±4.40	89.47±3.22	89.13±3.65
MDPNet (ours)	**93.28±1.54**	**93.26±4.00**	**93.65±3.65**	**93.42±3.70**
NEH (3 classes)				
ResNet50 [11]	96.59±0.18	96.33±0.28	96.83±0.17	96.58±0.22
DenseNet121 [12]	96.81±0.21	97.04±0.23	96.75±0.51	96.90±0.37
CrossViT [4]	90.98±0.37	90.29±0.44	91.61±0.42	90.95±0.43
Swin Transformer [14]	80.22±0.69	78.99±0.59	80.81±1.08	79.47±0.61
Deformable ProtoPNet [7]	97.18±0.39	97.38±0.45	96.88±0.38	97.13±0.41
MedMamba [23]	97.78±0.16	97.61±0.19	97.85±0.24	97.66±0.21
MDPNet (ours)	**97.90±0.37**	**97.71±0.41**	**98.02±0.34**	**97.87±0.38**
COVID-CT-Dataset (2 classes)				
ResNet50 [11]	71.95±2.98	71.87±2.94	71.91±2.98	71.89±2.96
DenseNet121 [12]	79.33±3.34	79.36±3.57	79.43±3.49	79.25±3.41
CrossViT [4]	75.97±4.36	75.78±4.78	76.37±4.17	76.07±4.46
MedMamba [23]	83.89±1.71	83.58±1.84	84.34±1.70	83.96±1.71
Swin Transformer [14]	85.37±2.90	85.42±2.80	85.57±2.87	85.49±2.80
Deformable ProtoPNet [7]	87.11±1.98	87.16±2.05	87.12±2.06	87.14±2.06
MDPNet (ours)	**87.65±2.54**	**87.74±2.65**	**87.72±2.42**	**87.73±2.49**

Table 2: Ablation studies (in %) on PVC and OCT datasets.

MSFEM	DPLM	LA	PVC Dataset		OCT Dataset	
			Accuracy	**F1**	**Accuracy**	**F1**
×	×	×	89.86±2.15	89.70±1.17	96.79±0.39	97.09±0.56
√	×	×	91.65±2.67	91.76±0.77	97.26±0.14	97.75±0.13
×	√	×	91.86±2.06	91.32±1.12	97.18±0.39	97.13±0.41
×	×	√	90.63±2.17	91.87±0.68	96.23±0.30	96.41±0.40
√	√	×	93.02±1.64	91.53±1.31	97.63±0.63	**98.19±0.40**
√	×	√	91.82±2.08	91.12±1.27	97.74±0.33	97.69±0.35
×	√	√	91.00±1.67	90.70±2.19	97.15±0.45	97.08±0.48
√	√	√	**93.28±1.54**	**92.59±2.27**	**97.90±0.37**	97.87±0.38

imbalance in PVC. We tune the temperature parameter τ for logit adjustment on the PVC dataset, as shown in Figure 3 (Left), and observe that the balanced error rate (BER) decreases as τ increases, reaching a minimum at τ=0.8, after which performance degrades, indicating over-calibration.

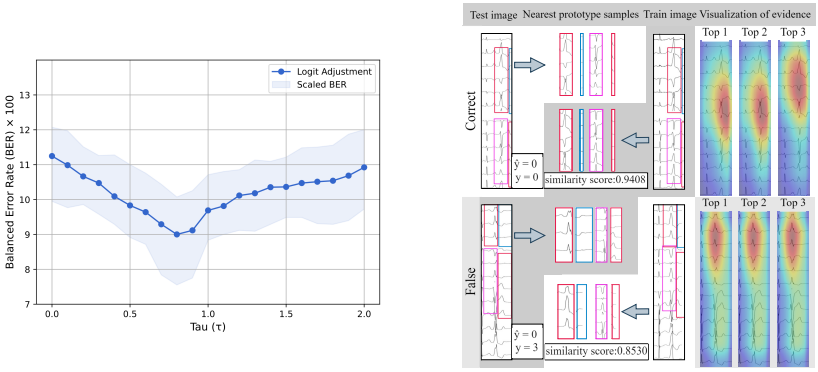

Fig. 3: (Left) Balanced Error Rate across different values of the temperature scaling parameter τ, with the lowest BER at $\tau = 0.8$. (Right) Correctly and incorrectly classified ECG images, each with corresponding prototypes.

3.4 Interpretability

Figure 3 (Right) illustrates the reasoning process of MDPNet on representative PVC test images. Classification predictions are supported by similarity scores to the nearest learned prototypes, providing case-based evidence in the embedding space. Despite strong performance on five-class PVC origin classification, we observe two representative failure modes: (1) **Prototype-feature mismatch under ambiguous morphology:** PVC patterns from adjacent origins may activate visually similar prototypes, yielding prototype matches that are less clinically aligned even when predictions are correct. Such cases typically arise when the available discriminative cues are weak or incomplete. (2) **Confusion under subtle inter-class differences:** overlapping QRS morphologies (e.g., PM vs. VA or LVOT vs. RVOT) can produce similar similarity scores across classes and lead to misclassification. These observations motivate anatomically sensitive prototype designs that better capture clinically meaningful differences.

4 CONCLUSION

In this paper, we propose MDPNet, a prototype-based framework that integrates a multi-scale spatial pyramid attention module, deformable prototype learning, and a logit-adjusted loss to improve both diagnostic accuracy and interpretability in medical image classification. MDPNet could support clinician-in-the-loop review by retrieving similar historical cases (prototypes) together with the evidence-bearing regions. Extensive experiments on three datasets demonstrate that MDPNet consistently outperforms strong baselines while providing prototype-grounded explanations, highlighting its potential for trustworthy computer-aided diagnosis. Nevertheless, the additional computational cost introduced by multi-scale attention, the uniform allocation of prototypes across

classes, and the abstract nature of prototype-based explanations for non-specialists remain important limitations. Future work will therefore focus on improving efficiency for deployment, developing adaptive prototype allocation strategies, and incorporating multimodal information (e.g., clinical text) to enhance explanation accessibility.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (62302329), Suzhou Planning Project of Science and Technology (SKY2023128), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2024B07), the Major Projects of Basic Science Research of Jiangsu Higher Education Institutions (25KJA520008), the Science and Technology Program of Jiangsu Province (BZ2024062), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Bai, X., Xia, Y.: Radioformer: Integrating radiologist inductive bias for tumor classification on multi-sequence mr images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 545–555 (2025). https://doi.org/10.1007/978-3-032-04927-8_52
2. Cai, W., He, J., Guo, X., Sun, H., Tong, H., Wei, B., Wu, H., Yao, D., Guo, D.: Regression-assisted classification for ct-based portal hypertension diagnosis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 533–542 (2025). https://doi.org/10.1007/978-3-032-05182-0_52
3. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This Looks Like That: Deep Learning for Interpretable Image Recognition. Advances in Neural Information Processing Systems (NeurIPS) pp. 8930–8941 (2019). https://doi.org/10.48550/arXiv.1806.10574
4. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). pp. 357–366 (2021). https://doi.org/10.1109/ICCV48922.2021.00041
5. Chen, J., et al.: Congenital Heart Disease Detection by Pediatric Electrocardiogram-Based Deep Learning Integrated With Human Concepts. Nature Communications **15**(1), 976 (2024). https://doi.org/10.1038/s41467-024-44930-y
6. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X.: Prior: Prototype Representation Joint Learning from Medical Images and Reports. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21361–21371 (2023). https://doi.org/10.1109/ICCV56702.2023.01952
7. Donnelly, J., Barnett, A.J., Chen, C.: Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10265–10275 (2022). https://doi.org/10.1109/CVPR52688.2022.01002

8. Gallée, L., Beer, M., Götz, M.: Interpretable Medical Image Classification Using Prototype Learning and Privileged Information. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 435–445 (2023). https://doi.org/10.1007/978-3-031-43895-0_41
9. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. The Lancet Digital Health **3**(11), e745–e750 (2021). https://doi.org/10.1016/S2589-7500(21)00208-9
10. He, H., Tang, J., Zhu, L., Li, K., Lu, Y.: Training-free test-time improvement for explainable medical image classification. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 638–648 (2025). https://doi.org/10.1007/978-3-032-05185-1_61
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 4700–4708 (2017). https://doi.org/10.1109/CVPR.2017.243
13. Kim, D., Seo, S., Kim, B., Choo, K., Jun, Y., Yun, M.: Prgnn: Pyramidal region graph neural network for region-based brain pet classification. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 554–563 (2025). https://doi.org/10.1007/978-3-032-05162-2_53
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
15. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-Tail Learning via Logit Adjustment. arXiv preprint arXiv:2007.07314 pp. 1–27 (2020). https://doi.org/10.48550/arXiv.2007.07314
16. Nasiri-Sarvi, A., Trinh, V.Q.H., Rivaz, H., Hosseini, M.S.: Vim4Path: Self-Supervised Vision Mamba for Histopathology Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6894–6903 (2024). https://doi.org/10.1109/CVPR52733.2024.00657
17. 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). https://doi.org/10.1038/s42256-019-0048-x
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-Cam: Visual Explanations from Deep Networks via Gradient-based Localization. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
19. Sotoudeh-Paima, S., Jodeiri, A., Hajizadeh, F., Soltanian-Zadeh, H.: Multi-Scale Convolutional Neural Network for Automated AMD Classification Using Retinal OCT Images. Computers in Biology and Medicine **144**, 105368 (2022). https://doi.org/10.1016/j.compbiomed.2022.105368
20. Wang, C., Liu, F., Chen, Y.g., Frazer, H., Carneiro, G.: Cross- and Intra-Image Prototypical Learning for Multi-Label Disease Diagnosis and Interpretation. IEEE Transactions on Medical Imaging **44**(6), 2568–2580 (2025). https://doi.org/10.1109/TMI.2024.3484545

21. Wang, J., Liu, H., Wang, X., Jing, L.: Interpretable Image Recognition by Constructing Transparent Embedding Space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 895–904 (2021). https://doi.org/10.1109/ICCV48922.2021.00094
22. Yu, Y., Zhang, Y., Cheng, Z., Song, Z., Tang, C.: Multi-Scale Spatial Pyramid Attention Mechanism for Image Recognition: An Effective Approach. Engineering Applications of Artificial Intelligence **133**, 108261 (2024). https://doi.org/10.1016/j.engappai.2024.108261
23. Yue, Y., Li, Z.: MedMamba: Vision Mamba for Medical Image Classification. arXiv preprint arXiv:2403.03849 pp. 1–26 (2024). https://doi.org/10.48550/arXiv.2403.03849
24. Zhang, Y., Li, H., Gao, Y., Duan, H., Huang, Y., Zheng, Y.: Prototype Correlation Matching and Class-Relation Reasoning for Few-Shot Medical Image Segmentation. IEEE Transactions on Medical Imaging **43**(11), 4041–4054 (2024). https://doi.org/10.1109/TMI.2024.3412420
25. Zhao, J., Zhang, Y., He, X., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020). https://doi.org/10.48550/arXiv.2003.13865