AGGRNet: Selective Feature Extraction and Aggregation for Enhanced Medical Image Classification

Ansh Makwe[1*], Akansh Agrawal[1*], Prateek Jain[1*], Akshan Agrawal[1,2*], and Priyanka Bagade[1✉]

[1] Indian Institute of Technology Kanpur, Kanpur, India
[2] Carnegie Mellon University, Pittsburgh, PA, USA

Corresponding author (✉): pbagade@iitk.ac.in

Abstract. Medical image analysis for complex tasks such as severity grading and disease subtype classification poses significant challenges due to intricate and similar visual patterns among classes, scarcity of labeled data, and variability in expert interpretations. Although deep learning models can capture complex visual patterns for medical image classification, many architectures still struggle to distinguish subtle classes because they do not adequately capture inter-class similarity and intra-class variability, which can lead to incorrect diagnoses. To address this, we propose the AGGRNet framework to extract informative and non-informative features to effectively understand fine-grained visual patterns and improve classification for complex medical image analysis tasks. Experimental results show that our model achieves state-of-the-art performance on 5 distinct medical imaging datasets, with the maximum improvement of 5% over SOTA models. The code is available at: `AGGRNet-Feature-Extraction-and-Aggregation`.

Keywords: Disease Subtype Classification · Severity Grading · Deep Learning.

1 Introduction

Medical image analysis plays a crucial role in accurate diagnosis and effective treatment planning. Despite significant advances in deep learning, medical image classification remains particularly challenging due to their complex visual characteristics. A fundamental difficulty arises from intra-class variability (the variability of features within the same class), and inter-class similarity (distinct classes share subtle yet overlapping features)[7]. For example, in ulcerative colitis, clinicians assign the Mayo Endoscopic Score (MES), ranging from 0 (inactive disease) to 3 (severe disease). The visual differences between adjacent grades can be subtle, while inflammation patterns within the same grade may vary considerably. Similarly, in datasets such as Kvasir [13], classification of disease subtypes,

[*] Equal contribution.

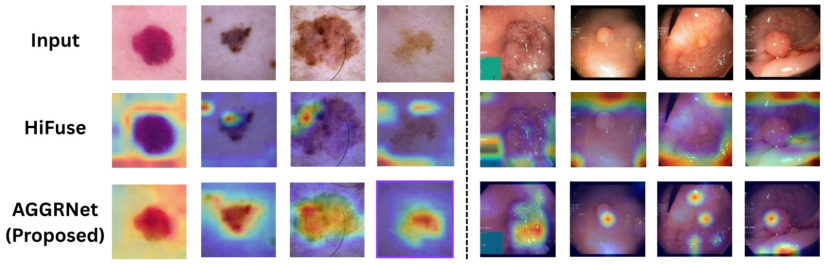

Fig. 1: Comparison of proposed AGGRNet framework with state-of-the-art HiFuse [7] architecture, showing improved identification of critical regions (Grad-CAM [16] visual results) on ISIC2018 (left to the dotted line) and Kvasir (right to the dotted line) datasets.

including polyps, esophagitis, ulcerative colitis, and anatomical landmarks, often involves categories with overlapping visual characteristics, further complicating reliable discrimination. Due to these intrinsic ambiguities, inconsistencies frequently arise in expert annotations [17][4].

The medical image classifiers have evolved from convolutional neural networks (CNNs) such as VGG [20], DenseNet [6], and ConvNeXt [10], to MLP-based models like MLP-Mixer [21], and Transformer-based models including ViT [3], DeiT [22], T2T-ViT [27], and Swin [9], which model global dependencies through self-attention but often lack strong local inductive bias. In order to extract both global semantics and local spatial features, recently multiple frameworks were proposed, such as class distance-weighted cross-entropy loss for ordinal MES prediction [15], hierarchical multi-scale feature fusion [7], XFMamba [28], i.e. a cross-fusion Mamba framework, that achieved better results.

Researchers often employ explainable AI techniques to validate medical image classifiers, aiming to demonstrate that the models attend to clinically relevant regions of the image. However, this practice implicitly assumes that deep learning models are capable of automatically extracting all the critical features necessary for accurate prediction. As illustrated in Figure 1, the model may not always focus on the most diagnostically relevant regions when generating predictions. Given the challenges of capturing both inter-class similarity and intra-class variability in medical imaging, it is essential to guide the model toward learning meaningful and discriminative features during training.

To address these limitations, we propose the AGGRNet framework, motivated by two key observations. First, medical images typically contain fine-grained pathological cues embedded within complex backgrounds, making it essential to separate diagnostically relevant features (hereafter referred to as informative features) from background or redundant information (non-informative features). By explicitly disentangling relevant and irrelevant features, the model can identify visual patterns that consistently contribute to the final classification decision, thereby reducing the adverse impact of intra-class variability. Second,

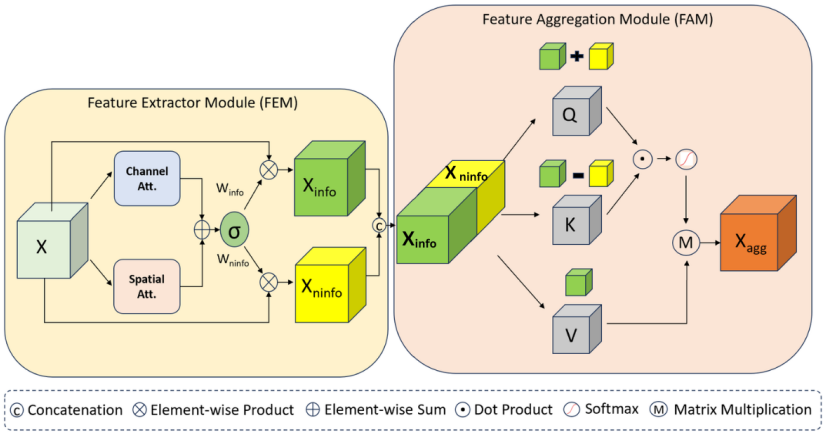

Fig. 2: The proposed Feature Extraction and Aggregation (FEA) Module

many anatomical and pathological distinctions cannot be resolved solely through local features but instead require an understanding of global spatial relationships [7]. Thus, capturing long-range dependencies in medical images facilitates better differentiation between classes exhibiting high inter-class similarity.

Building upon this, AGGRNet framework incorporates a novel Feature Extraction and Aggregation (FEA) module to separate informative and non informative features and capture global dependencies. With the FEA module, the proposed AGGRNet framework captures intra-class variability and inter-class similarity that results in outperforming existing state-of-the-art(SOTA) results on distinct medical imaging datasets.

The main contributions can be summarized as follows: (1) We propose a novel generalized AGGRNet framework that provides a unified single-model solution for both severity grading and disease subtype classification. (2) As a part of this framework, we propose a novel FEA module, for extracting informative and non-informative features, while capturing global spatial dependencies. (3) The AGGRNet framework consistently outperforms SOTA models across distinct medical imaging datasets.

2 Methodology

This section presents details of Feature Extraction and Aggregation Module (FEA) and the overall proposed AGGRNet framework.

2.1 Feature Extraction and Aggregation Module

The proposed FEA module (Figure 2) consists of two components: (1) FEM and (2) FAM. The **FEM** addresses the critical challenge of distinguishing between diagnostically relevant features and background in medical images. Given an

input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where H, W, and C represent height, width, and number of channels respectively. The proposed FEM decomposes $\mathbf{X}$ into two components: informative features $\mathbf{X}_{\text{info}}$ and non-informative features $\mathbf{X}_{\text{ninfo}}$. In order to identify $\mathbf{X}_{\text{info}}$ and $\mathbf{X}_{\text{ninfo}}$, the FEM module employs spatial and channel attention mechanisms [24] to generate attention weights. The spatial attention layer $\text{SA}(\cdot)$ focuses on identifying spatially significant regions, while the channel attention layer $\text{CA}(\cdot)$ emphasizes important feature channels. The outputs of both attention layers are added element-wise and passed through a sigmoid function to produce normalized attention scores.

$$\mathbf{SA_{out}} = \text{SA}(\mathbf{X}), \quad \mathbf{CA_{out}} = \text{CA}(\mathbf{X}), \quad \mathbf{S} = \sigma(\mathbf{SA_{out}} + \mathbf{CA_{out}}) \tag{1}$$

We introduce an adaptive threshold parameter τ to segregate informative and non-informative features based on the attention score matrix $\mathbf{S}$. Initialized at 0.5, τ is learned during training and generates two complementary binary weight matrices, $\mathbf{W}_{\text{info}}$ and $\mathbf{W}_{\text{ninfo}}$, which are constructed using the rule:

$$\mathbf{W}_{\text{info}}[i,j,k] = \begin{cases} 1 & \text{if } \mathbf{S}[i,j,k] \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad \mathbf{W}_{\text{ninfo}}[i,j,k] = \begin{cases} 1 & \text{if } \mathbf{S}[i,j,k] < \tau \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{S}[i,j,k]$ represents the attention score at location (i,j) and channel k. These binary matrices act as selective masks, with $\mathbf{W}_{\text{info}}$ identifying informative regions, while $\mathbf{W}_{\text{ninfo}}$ captures the complementary non-informative region. The segregated features are obtained through element-wise multiplication as:

$$\mathbf{X}_{\text{info}} = \mathbf{W}_{\text{info}} \odot \mathbf{X}, \quad \mathbf{X}_{\text{ninfo}} = \mathbf{W}_{\text{ninfo}} \odot \mathbf{X} \tag{3}$$

The **FAM** leverages novel cross-attention mechanisms to enable informative features to attend to non-informative features, boosting the model's capability to capture global-level contextual relationships. Here, the Query (Q), Key (K), and Value (V) are defined to maximize the focus on the informative features as:

$$\mathbf{Q} = \mathbf{X}_{\text{info}} + \mathbf{X}_{\text{ninfo}}, \quad \mathbf{K} = \mathbf{X}_{\text{info}} - \mathbf{X}_{\text{ninfo}}, \quad \mathbf{V} = \mathbf{X}_{\text{info}} \tag{4}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

where d_k represents the dimension of the key vector, and the softmax function ensures normalization of attention weights.

Mathematical Justification for Q-K Formulation The mathematical formulation of $\mathbf{Q}$ and $\mathbf{K}$ is designed to create a bias toward informative features $\mathbf{X}_{\text{info}}$. When computing attention weights through $\mathbf{Q}\mathbf{K}^T$, our formulation yields:

$$\begin{aligned} \mathbf{Q}\mathbf{K}^T &= (\mathbf{X}_{\text{info}} + \mathbf{X}_{\text{ninfo}})(\mathbf{X}_{\text{info}} - \mathbf{X}_{\text{ninfo}})^T \\ &= \mathbf{X}_{\text{info}}\mathbf{X}_{\text{info}}^T - \mathbf{X}_{\text{ninfo}}\mathbf{X}_{\text{ninfo}}^T + (\mathbf{X}_{\text{ninfo}}\mathbf{X}_{\text{info}}^T - \mathbf{X}_{\text{info}}\mathbf{X}_{\text{ninfo}}^T) \end{aligned} \tag{6}$$

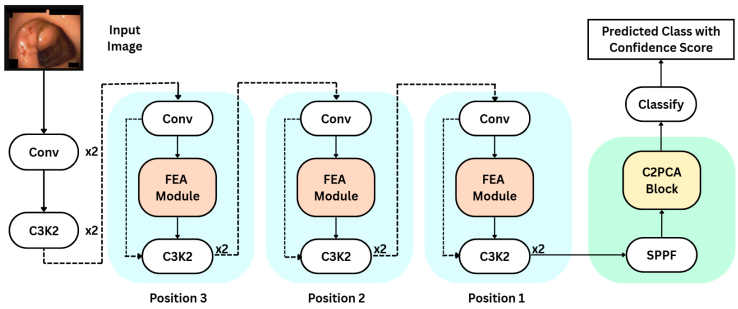

Fig. 3: The proposed architecture AGGRNet with the novel FEA module and C2PCA block.

Equation 6 reveals three key components that create the informative bias:
Positive Bias Term for Informative Regions: $\mathbf{X}_{\text{info}}\mathbf{X}_{\text{info}}^T$ acts like a positive self-similarity among informative features, encouraging high attention around informative regions.
Negative Bias Term against Non-Informative Regions: $-\mathbf{X}_{\text{ninfo}}\mathbf{X}_{\text{ninfo}}^T$ acts like negative self-similarity among non-informative features and suppresses attention around non-informative regions.
Cross-Modal Interaction Terms: The $(\mathbf{X}_{\text{ninfo}}\mathbf{X}_{\text{info}}^T - \mathbf{X}_{\text{info}}\mathbf{X}_{\text{ninfo}}^T)$ terms capture asymmetric interactions between informative and non-informative parts, allowing subtle contextual modulation.

Therefore, this formulation establishes a **contrast-based cross-attention mechanism**, and also aligns with attention mechanism principles where similarity between queries and keys determines relevance. The final aggregated features incorporate global contextual information.

2.2 Proposed Architecture - AGGRNet

Figure 3 illustrates the proposed AGGRNet architecture with FEA modules. The FEA module is architecture-agnostic and can be integrated into any CNN architecture. For the proposed AGGRNet architecture, we integrate the FEA module into the YOLOv11 classification backbone. Based on CNN feature hierarchies, we propose positioning FEA in deeper layers, where task-specific rich features are available. To preserve feature information and ensure stable gradient flow, a residual connection is used.
C2PCA Module: The built-in C2PSA module in YOLOv11 utilizes self-attention mechanisms to capture global semantics. Since the FEA module already extracts globally attended informative features, we replace the self-attention component with a channel attention module, creating the C2PCA (Cross Stage Partial with Channel Attention) block to encode anatomical patterns across feature channels.

Table 1: Performance comparison on LIMUC, Path & RetinaMNIST datasets.

(a) LIMUC Dataset

Methods	LIMUC			
	Acc(%)	Macro-f1	QWK	MAE
Cross-entropy with Inception-v3 [15]	76.0	68.3	83.6	0.253
CORN with Inception-v3 [18]	76.0	68.3	84.3	0.25
CO2 with Inception-v3 [1]	76.5	68.5	84.8	0.24
HO2 with Inception-v3 [1]	76.6	69.0	84.6	0.242
CDW-CE with Inception-v3 [15]	78.8	72.6	86.8	0.215
SATOMIL [19]	69.0	67.4	82.6	-
Proposed AGGRNet	**81.0**	**73.9**	**88.1**	**0.194**

(b) Path & RetinaMNIST Datasets

Methods	PathMNIST		RetinaMNIST	
	AUC	Acc(%)	AUC	Acc(%)
ResNet-18 (28) [5]	98.3	90.7	71.7	52.4
ResNet-18 (224) [5]	98.9	90.9	71.0	49.3
ResNet-50 (28) [5]	99.0	91.1	72.6	52.8
ResNet-50 (224) [5]	98.9	89.2	71.6	51.1
auto-sklearn [26]	93.4	71.6	69.0	51.5
AutoKeras [26]	95.9	83.4	71.9	50.3
Google AutoML Vision [26]	94.4	72.8	75.0	53.1
Proposed AGGRNet	**99.6**	**92.6**	**74.5**	**53.7**

3 Experiments

3.1 Datasets

We evaluated AGGRNet on multiple public datasets covering severity grading and disease subtype classification tasks. **LIMUC** [14] consists of 11,276 endoscopic images from 564 patients, annotated with Mayo Endoscopic Score (MES 0–3). **ISIC2018** [2,23] consists of 10,015 dermoscopic images across seven skin lesion classes. **Kvasir** [13] consists of 4,000 gastrointestinal endoscopic images across eight categories. **PathMNIST** [26] consists of 107,180 histopathology images categorized into nine tissue classes. **RetinaMNIST** [26] consists of 1,600 retinal images labeled into five diabetic retinopathy severity grades.

3.2 Implementation Details

AGGRNet was implemented in PyTorch and trained on an NVIDIA RTX A6000 GPU. The YOLOv11 backbone was initialized with ImageNet pretrained weights. Images were resized to 224×224, and training used SGD with a learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}. For reproducibility, we used identical initial random weights for the AGGRNet across runs in the same hardware and software environment.

3.3 Evaluation Metrics

The performance of AGGRNet was evaluated using Accuracy (Acc), Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Precision (P), Recall (R), F1-score (F1), and Area under the Curve (AUC), capturing various aspects of classification efficacy. Metric selection for each dataset is guided by results reported in the literature on SOTA models.

3.4 Results

Tables 1a–1b and 2 demonstrate consistent improvement of AGGRNet over SOTA methods on five distinct medical imaging datasets.

Image Label	Mayo 0	Mayo 1	Mayo 2	Mayo 3
Images → / Model ↓				
CDW with InceptionV3	0.6845	0.4981	0.4661	0.0
AGGRNet (Ours)	0.9678	0.9137	0.7907	0.7098

(a) LIMUC dataset: CDW-CE with Inception-V3 vs proposed AGGRNet

Image Label	Normal-Cecum	Dyed-Lifted-Polyps	Normal-Z-Line
Images → / Model ↓			
HiFuse	0.7637	0.5235	0.6877
AGGRNet (Ours)	0.9844	0.999	0.8734

(b) Kvasir dataset: HiFuse vs proposed AGGRNet

Fig. 4: Class-wise predicted probability comparison between state-of-the-art models and the proposed AGGRNet on LIMUC and Kvasir datasets.

Results on LIMUC dataset: AGGRNet achieves a 2.2% gain in Acc, 1.3% in F1, 1.3% in QWK, and reduces MAE by 0.021 compared to the SOTA model (Table 1a). Figure 4a shows that AGGRNet produces higher confidence for correct Mayo classes and better distinguishes adjacent severity grades, reflected in the superior QWK score.
Results on PathMNIST and RetinaMNIST datasets: In PathMNIST, different tissue states can look locally similar, and the same class may vary across patients and slide regions, making appearance-based classification ambiguous. The higher AUC (Table 1b) of AGGRNet indicates better class discrimination. RetinaMNIST is particularly challenging due to its tiny image size, which limits fine-grained visual details; despite this, it achieves higher Acc (Table 1b).
Results on ISIC2018 dataset: AGGRNet improves Acc by 1.25%, F1 by 1.78%, and P by 5.73% compared to SOTA models (Table 2). Notably, AGGRNet outperforms Transformer-based models, Swin-B and Conformer as evidenced by enhanced precision.
Results on Kvasir dataset: AGGRNet achieves gains of 5.48% in Acc, 5.47% in F1, 5.75% in P, and 5.47% in R over SOTA methods (Table 2). It outperforms hybrid CNN-Transformer models such as Conformer and BiFormer, with improved recall for categories like polyps and esophagitis. Figure 4b further shows consistently higher class-wise confidence compared to HiFuse.

3.5 Ablation Study

Table 3 shows a detailed ablation study of the proposed AGGRNet by systematically assessing the contribution of each architectural component. We start by evaluating a plain ResNet18 model and then added FEA block between feature layers 4 and 5 of the ResNet18 architecture. The addition of the FEA block improved the classification results. Then we tried YOLOv11 classification backbone with the built-in C2PSA module, and then replaced C2PSA with C2PCA to study the effect of our improved attention mechanism. Since YOLOv11 per-

Table 2: Performance comparison on ISIC2018 and Kvasir datasets.

Methods	ISIC2018					Kvasir			
	Params(M)	Acc(%)	Macro-F1	Prec	Rec	Acc(%)	Macro-F1	Prec	Rec
VGG-19 [20]	143.68	79.2	61.8	63.7	60.8	77.7	77.7	77.8	77.8
Mixer-L/16 [21]	208.20	78.9	59.8	61.4	59.2	74.3	74.1	74.4	74.3
T2T-ViT_t-24 [27]	64.00	77.6	57.2	59.6	55.9	76.9	76.9	77.6	76.9
DeiT-base [22]	86.57	72.3	41.0	47.2	44.1	52.1	48.5	56.7	52.3
ViT-B/16 [3]	86.86	78.3	60.9	64.2	60.5	76.1	75.9	76.5	76.2
ViT-B/32 [3]	88.30	77.9	57.5	58.7	56.9	73.8	73.5	74.2	73.7
Swin-B [9]	87.77	79.8	63.9	65.1	63.6	77.3	77.3	77.7	77.4
Conformer-base-p16 [12]	83.29	82.7	72.4	73.3	71.7	84.2	84.3	84.4	84.4
ConvNeXt-B [10]	88.59	79.9	63.2	64.9	62.1	74.6	74.4	75.7	74.6
PerViT-M [11]	43.04	81.6	67.7	68.2	67.3	82.4	82.3	82.9	82.4
Focal-B [25]	87.10	79.6	62.9	65.7	60.7	78.0	77.9	78.2	78.0
UniFormer-B [8]	50.02	82.4	68.4	70.7	66.5	83.1	83.0	83.1	83.1
BiFormer-B [29]	56.04	82.7	68.9	72.7	66.5	84.2	84.3	84.7	84.2
HiFuse-Tiny [7]	82.49	82.9	72.9	73.7	72.9	84.8	84.9	84.9	84.9
HiFuse-Small [7]	93.82	83.6	72.7	72.7	73.1	86.1	86.1	86.2	86.1
HiFuse-Base [7]	127.80	85.8	75.3	74.6	76.6	85.9	86.1	86.3	86.0
Proposed AGGRNet	**38.65**	**87.1**	**77.1**	**80.3**	**74.8**	**91.6**	**91.6**	**92.0**	**91.6**

Table 3: Ablation study of C2PCA, FEA modules, and SPPF block.

Architecture	**Acc(%)**
ResNet18	75.6
ResNet18 + FEA	77.1
YOLOv11 (with default C2PSA)	77.5
YOLOv11 backbone + C2PCA	79.3
YOLOv11 backbone + C2PCA + FEA@1	79.1
YOLOv11 backbone + C2PCA + FEA@1,2	79.3
YOLOv11 backbone + C2PCA + FEA@1,2,3	80.7
Proposed AGGRNet (YOLOv11 backbone + C2PCA + FEA@1,2,3 + SPPF)	**81.0**

formed better than ResNet18, we used it as the backbone for all the experiments. Replacing C2PSA with the C2PCA block increases the accuracy by 2%. We progressively insert the proposed FEA modules at three hierarchical positions starting from deeper layers within the backbone, as highlighted in the architecture diagram, Figure 3. FEA@1, FEA@1,2 and FEA@1,2,3 represent addition of FEA modules at position 1, 1 and 2, and 1, 2 and 3 respectively. The performance improves steadily with each additional FEA module. Finally, integrating the SPPF block into AGGRNet improves accuracy, highlighting the benefit of combining SPPF with hierarchical feature aggregation.

4 Conclusion

In this paper, we propose a novel framework, AGGRNet, with the capability to capture both informative and non-informative features using the proposed Feature Extraction and Aggregation (FEA) module for effectively classifying disease subtypes and severity. We validated the proposed AGGRNet framework through

extensive experiments on distinct datasets, while improving the medical image classification performance over the state-of-the-art models. This approach can assist clinicians to diagnose and treat patients more accurately, reducing reliance on manual scoring methods that are often prone to subjective ambiguity.

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

References

1. Albuquerque, T., Cruz, R., Cardoso, J.: Ordinal losses for classification of cervical cancer risk. PeerJ Comput. Sci. **7**, e457 (2021)
2. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018. arXiv preprint arXiv:1902.03368 (2018)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (ICLR) (2021)
4. Hashash, J., et al.: Inter- and intraobserver variability on endoscopic scoring systems in crohn's disease and ulcerative colitis: A systematic review and meta-analysis. Inflamm. Bowel Dis. **30**(11), 2217–2226 (2024)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 770–778 (2016)
6. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.: Densely connected convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4700–4708 (2017)
7. Huo, X., et al.: Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomed. Signal Process. Control **87**, 105534 (2024)
8. Li, K., et al.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2023). https://doi.org/10.1109/TPAMI.2023.3282631
9. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 9992–10002 (2021)
10. Liu, Z., et al.: A convnet for the 2020s. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11966–11976 (2022)
11. Min, J., Zhao, Y., Luo, C., Cho, M.: Peripheral vision transformer. arXiv preprint arXiv:2206.06801 (2022)
12. Peng, Z., et al.: Conformer: Local features coupling global representations for visual recognition. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 357–366 (2021)
13. Pogorelov, K., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: ACM Multimedia Syst. Conf. (MMSys). pp. 164–169 (2017)
14. Polat, G., et al.: Labeled images for ulcerative colitis (limuc) dataset. Zenodo (2022). https://doi.org/10.5281/zenodo.5827695
15. Polat, G., Ergenc, I., Kani, H., Alahdab, Y., Atug, O., Temizel, A.: Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. arXiv preprint arXiv:2202.05167 (2022)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019).

https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7
17. Sharara, A., Malaeb, M., Lenfant, M., Ferrante, M.: Assessment of endoscopic disease activity in ulcerative colitis: Is simplicity the ultimate sophistication? Inflamm. Intest. Dis. **7**(1), 7–12 (2022)
18. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Anal. Appl. **26**(3), 941–955 (2023)
19. Shiku, K., Nishimura, K., Suehiro, D., Tanaka, K., Bise, R.: Ordinal multiple-instance learning for ulcerative colitis severity estimation with selective aggregated transformer. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV). pp. 4290–4299 (2025)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Tolstikhin, I., et al.: Mlp-mixer: An all-mlp architecture for vision. In: Adv. Neural Inf. Process. Syst. (NeurIPS) (2021)
22. Touvron, H., et al.: Training data-efficient image transformers & distillation through attention. In: Proc. Int. Conf. Mach. Learn. (ICML). vol. 139, pp. 10347–10357 (2021)
23. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018)
24. Woo, S., Park, J., Lee, J.Y., Kweon, I.: Cbam: Convolutional block attention module. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 3–19 (2018)
25. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. In: Adv. Neural Inf. Process. Syst. (NeurIPS) (2022)
26. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. Sci. Data **10**(1), 1 (2023)
27. Yuan, L., et al.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 538–547 (2021)
28. Zheng, X., Chen, X., Gong, S., Griffin, X., Slabaugh, G.: Xfmamba: Cross-fusion mamba for multi-view medical image classification. In: Med. Image Comput. Comput.-Assist. Interv. (MICCAI). pp. 672–682. Springer (2025)
29. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.: Biformer: Vision transformer with bi-level routing attention. arXiv preprint arXiv:2303.08810 (2023)