# Med-Former: A Transformer based Architecture for Medical Image Classification

G. Jignesh Chowdary, Zhaozheng Yin

Stony Brook University, NY, USA.

**Abstract.** In recent years, transformer-based image classification methods have demonstrated remarkable effectiveness across various image classification tasks. However, their application to medical images presents challenges, especially in the feature extraction capability of the network. Additionally, these models often struggle with the efficient propagation of essential information throughout the network, hindering their performance in medical imaging tasks. To overcome these challenges, we introduce a novel framework comprising Local-Global Transformer module and Spatial Attention Fusion module, collectively referred to as Med-Former. These modules are specifically designed to enhance the feature extraction capability at both local and global levels and improve the propagation of vital information within the network. To evaluate the efficacy of our proposed Med-Former framework, we conducted experiments on three publicly available medical image datasets: NIH Chest X-ray14, DermaMNIST, and BloodMNIST. Our results demonstrate that Med-Former outperforms state-of-the-art approaches underscoring its superior generalization capability and effectiveness in medical image classification. Code: `https://github.com/jignesh9999/class`

**Keywords:** Medical Image Classification · Transformers · Computer Aided Diagnosis · Local-global Feature Extraction · Spatial Attention Fusion

## 1 Introduction

Medical image classification is pivotal in the development of Computer-Assisted Diagnostic (CAD) systems, reducing diagnosis time and assisting in diagnosis [1,18,2]. However, medical image classification presents challenges due to the intrinsic complexities of diseases, such as very small infected regions (e.g., nodules in chest x-rays), poor contrast between background and infected regions, and diseased areas resembling other normal areas (e.g., diseased black dots on skin similar to mole marks).

Recent advancements in deep learning [16] have led to the widespread adoption of convolutional neural network (CNN)-based approaches for natural image recognition tasks. Despite their remarkable performance, CNNs have inherent limitations. For instance, each convolutional kernel can only focus on a sub-region of the input image due to its inherent inductive biases, complicating the extraction of global contextual information crucial for medical image classification.

To tackle this challenge, researchers introduced Inception networks [17], capable of extracting multi-scale information. However, these networks encounter issues such as vanishing gradients and information loss from earlier layers. To address these concerns, researchers developed Residual networks [7] and DenseNets [8]. Residual networks incorporate a residual (or skip) connection between the input and output of each convolutional block, while DenseNets utilize dense connections between all layers, with each layer's input being the concatenated output of all preceding layers. Although these networks capture information from earlier layers, they may not enable the model to focus attentions on specific regions essential for medical image classification, as they lack attention mechanisms to emphasize important features.

Recently, Transformer-based approaches with self-attention mechanisms have been developed for image recognition, such as Vision Transformers (ViT) [3], capable of capturing better contextual information compared to CNNs [10,6]. These methods partition the input image into non-overlapping patches and utilize a window (a collection of patches) for self-attention computation. To further enhance contextual information extraction, researchers introduced Swin Transformers [12]. These networks employ sequentially connected two transformer blocks with different window strategies for computing self-attention. However, these networks do not fully capture information at local and global levels and suffer from information loss from earlier layers.

To address these limitations, we introduce Med-Former, a transformer-based approach adept at enhancing the capability of extracting essential information at both local and global levels while mitigating issues of information loss during the propagation of essential information throughout various layers of the network. Our contributions are outlined as follows:

- We introduce a Local-Global Transformer (LGT) module, inspired by the structure of the Swin Transformer module. This module comprises two parallel attention computation paths, each with different window sizes in both the window and shifted window blocks. This design enhances the extraction of contextual information at both local and global levels, thereby improving the performance of medical image classification.
- We propose a Spatial Attention Fusion (SAF) module designed to fuse information from earlier layers and facilitate the propagation of more essential information through the network.
- Based on the LGT and SAF modules, our Med-Former has demonstrated superior performance compared to the latest state-of-the-art on benchmark datasets with various imaging modalities and diseases, including thoracic disease classification from chest X-rays, skin lesion classification from dermoscopic images, and blood cell classification from microscopic images, justifying Med-Former's effectiveness in generalizing to different medical image classification tasks.

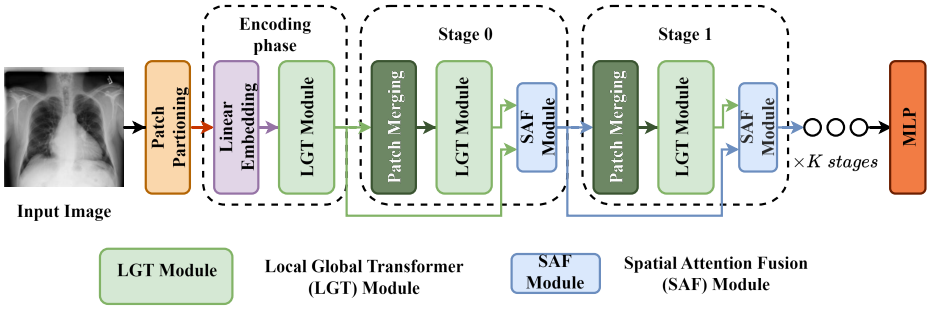


**Fig. 1.** Architecture of the proposed Med-Former.

## 2 Methodology

In this section, first we describe the overview of Med-Former architecture. Then we introduce the technical details of the newly designed Local-Global Transformer (LGT) module, and Spatial Attention Fusion (SAF) module, which are key elements of the Med-Former.

### 2.1 Overview

Figure 1 illustrates the Med-Former architecture, comprising a Patch Partitioning layer, a Linear Embedding layer, LGT modules, Patch Merging layers, SAF modules, and an MLP for classification. The Patch Partitioning layer divides the input image into patches of size $s \times s$, where $s$ represents the patch width or height. These patches are processed via a Linear Embedding layer in the Embedding phase before being processed by a LGT module and forwarded to subsequent stages. The subsequent stages incorporate Patch Merging layers, LGT modules, and SAF modules. Each stage utilizes a Patch Merging layer to downsample the input feature maps. A SAF module is employed to fuse outputs from both the previous stage's SAF module and the current stage's LGT module, improving the continuity of essential feature propagation and reducing information loss. Stage 0 differs from the $K$ sequential stages slightly, which incorporates fused outputs from the Embedding phase's LGT module and the current stage's LGT module, thereby enhancing the contextual understanding of the input image.

### 2.2 Local-Global Transformer (LGT) Module

The proposed LGT module diverges from the conventional Swin-Transformer architecture by integrating two parallel paths (Figure 2(a)): the Global path $G_p$ and the Local path $L_p$ for computing Multihead Self-attention (MSA) with varying window sizes in both of the two transformer blocks (i.e., block $l$ and $l+1$). The $G_p$ employs a global window of size $m \times m$, and $L_p$ employs a local window of size $n \times n$, where $m > n$. This configuration facilitates the extraction

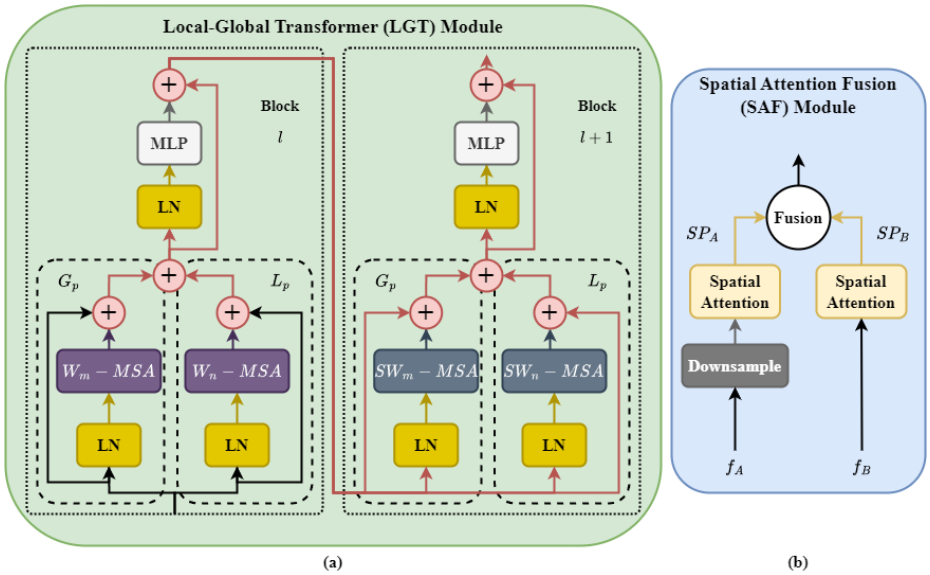


**Fig. 2.** Architecture of the proposed Local-Global Transformer (LGT) module (a) and Spatial Attention Fusion (SAF) module (b). LN represents Layer Normalization, W-MSA represents Window-Multihead Self-Attention, SW-MSA represents Shifted Window-Multihead Self-Attention. The subscripts $m$ and $n$ in W-MSA and SW-MSA represent the window sizes. MLP represents Multi-Layer Perceptron. $SP_A$ and $SP_B$ represent the spatial attention maps of feature maps $f_A$ and $f_B$, respectively.

of both global and local information at the window level, thereby enhancing the feature representation learning and classification performance. The outputs of $G_p$ and $L_p$ in block $l$ with MSA on various windows, represented as $W_m - MSA$ and $W_n - MSA$, are later combined and propagated to the next block $l+1$ with MSA on shifted windows, denoted as $SW_m - MSA$ and $SW_n - MSA$, respectively.

### 2.3 Spatial Attention Fusion (SAF) Module

The proposed SAF Module (Figure 2(b)) is utilized to fuse feature maps from the preceding layers and stages, facilitating the transfer of crucial information within the network with less information loss. This module accepts two feature maps, $f_A$ and $f_B$, where $f_A$ is the feature map from the preceding stage, and $f_B$ is the feature map from the previous layer. Initially, it downsamples the feature map $f_A$ to match the dimensions of the feature map $f_B$. Subsequently, spatial attention maps $SP_A$ and $SP_B$ are computed for feature maps $f_A$ and $f_B$, respectively. Finally, the fused output of $SP_A$ and $SP_B$ is forwarded to the succeeding stage.

## 3 Results

In this section, we first describe the datasets, evaluation metrics, and implementation details. Then, we present the comparison with latest state-of-the-art methods, and discuss the ablation results.

### 3.1 Datasets and Performance Metrics

To assess the generalization capability of the proposed Med-Former, it undergoes testing on three distinct medical image classification tasks, each representing different imaging modalities and disease types: thoracic disease classification from chest X-rays (NIH Chest X-ray14, denoted as ChestX) [19]; skin lesion classification from dermoscopic images (DermaMNIST, denoted as DM) [20]; and blood cell classification from microscopic images (BloodMNIST, denoted as BM) [20].

The study employs the official dataset splits for training and evaluating Med-Former. For ChestX, the training/validation set consists of 86,524 images, while the test set comprises 25,596 images. For DM, the training/validation set includes 8,010 images, and the test set contains 2,005 images. For BM, the training/validation set consists of 13,671 images, and the test set comprises 3,421 images.

Following other state-of-the-art (SOTA) methods [9,14,13,15,4,5,11], we employed classification accuracy (ACC) and the area under the curve (AUC) for the evaluation on the DM and BM datasets, while the AUC was used for the ChestX dataset.

### 3.2 Implementation Details

For the three datasets, the number of stages of Med-Former is determined to be $K = 3$, by cross-validation. The model is trained by minimizing the Cross-Entropy loss for 400 epochs, using a batch size of 16 and an initial learning rate of 0.001. Additionally, the learning rate is decayed by a factor of 0.1 every 100 epochs. All experiments are conducted on an NVIDIA Tesla V100 GPU with 32 GB of RAM.

**Table 1.** Performance comparison with transformer-based approaches. The best performance is highlighted in bold.

| Method | Datasets | | | | |
|---|---|---|---|---|---|
| | NIH | DM | | BM | |
| | AUC | ACC | AUC | ACC | AUC |
| Vision Transformer (ViT) [3] | 0.836 | 0.739 | 0.883 | 0.921 | 0.985 |
| Swin Transformer [12] | 0.841 | 0.753 | 0.903 | 0.935 | 0.991 |
| Ours | **0.876** | **0.783** | **0.946** | **0.965** | **0.997** |

### 3.3 Performance Comparison

Firstly, we evaluate the performance of the proposed Med-Former against existing transformer-based approaches, namely Vision Transformer (ViT) [3] and Swin Transformer [12]. To ensure a fair comparison, ViT and Swin Transformer are trained and evaluated on the ChestX, DM, and BM datasets using the same evaluation protocol. The results of this comparison are summarized in Table 1. As shown, the proposed Med-Former outperforms ViT and Swin Transformer, showcasing its superior generalization capability.

Secondly, we compare Med-Former against SOTA approaches in Table 2. All these approaches have reported their results using the same evaluation protocol as our work, enabling a fair comparison. As observed, Med-Former surpasses the SOTA methods, emphasizing the effectiveness of its local-global feature learning and the information propagation through the network.

Lastly, we illustrate the performance of Med-Former through some correctly classified and misclassified samples in Figure 3. The misclassifications arise from diseases sharing similar characteristics, such as white lung fields in chest X-rays, widespread infection (not as a cluster) in dermoscopic images, and similar extracellular and intracellular structures in microscopic images. In the future, to overcome these limitations, we plan to enrich our model with additional information, such as patient symptoms, to aid in diagnosis.

**Table 2.** Comparison with state-of-the-art approaches on ChestX [19], DM [20], and BM [20] datasets.

| ChestX [19] | | DM [20] | | | BM [20] | | |
|---|---|---|---|---|---|---|---|
| Method | AUC | Method | ACC | AUC | Method | ACC | AUC |
| Kamal et al. [9] | 0.850 | MedViT [14] | 0.773 | 0.920 | MedViT [14] | 0.954 | 0.996 |
| Luo et al. [13] | 0.834 | BP-CapsNet [11] | 0.774 | 0.923 | BP-CapsNet [11] | 0.946 | 0.996 |
| Nie et al. [15] | 0.857 | SADAE [4] | 0.759 | 0.927 | | | |
| DGFN [5] | 0.850 | | | | | | |
| Ours | **0.876** | Ours | **0.783** | **0.946** | Ours | **0.965** | **0.997** |

### 3.4 Ablation results

We conducted a series of ablation experiments to assess the effectiveness of various modules within Med-Former. The quantitative results of these experiments are summarized in Table 3, and the qualitative comparison is presented in Figure 4.

- In row 2, we concatenated the information from the two preceding stages in the original Swin Transformer and passed them as input to the current stage. This resulted in improved performance compared to row 1, indicating that information flow from earlier layers enhances classification performance.

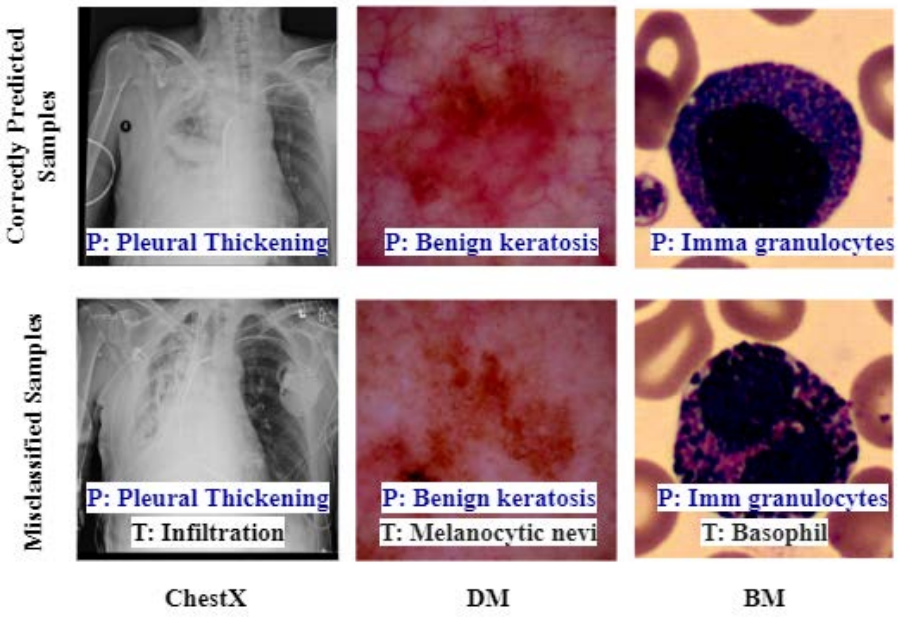


**Fig. 3.** Correctly classified, and misclassified samples by the Med-Former. P - Predicted class; T - Ground truth class.

- Row 3 builds upon row 1 by passing the combined output of the preceding stages using the Spatial Attention Fusion (SAF) module. This further improves performance, suggesting that the SAF module facilitates the flow of important information from earlier stages.
- In row 4, we present the performance of Med-Former with Local-Global Transformer (LGT) modules but without information from the preceding layers or SAF modules. The enhanced performance of row 4 compared to the original Swin Transformer in row 1 underscores the local and global feature extraction capability of the LGT module for improving medical image classification.
- Row 5 extends row 4 by adding information from the preceding layers and stages and fusing the features using the standard concatenation. This further enhances the classification performance compared to row 4, indicating the effective propagation of local and global information through the network.
- Finally, in row 6, we present the performance of the complete Med-Former model. Compared to other configurations, this model achieves the highest performance, highlighting the local-global feature extraction capability of LGT modules and the effective flow of essential information through the network facilitated by SAF modules.

**Table 3.** Ablation experiments were conducted on the ChestX [19], DM [20], and BM [20] datasets. The 'IF' denotes the information flow from earlier layers.

| Row | Swin Transformer | LGT | IF with Concat | IF with SAF | Datasets | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | NIH | DM | | BM | |
| | | | | | AUC | ACC | AUC | ACC | AUC |
| 1 | ✓ | | | | 0.841 | 0.753 | 0.903 | 0.935 | 0.991 |
| 2 | ✓ | | ✓ | | 0.843 | 0.760 | 0.909 | 0.938 | 0.991 |
| 3 | ✓ | | | ✓ | 0.846 | 0.763 | 0.914 | 0.942 | 0.992 |
| 4 | | ✓ | | | 0.848 | 0.761 | 0.912 | 0.941 | 0.992 |
| 5 | | ✓ | ✓ | | 0.852 | 0.778 | 0.925 | 0.956 | 0.994 |
| 6 | | ✓ | | ✓ | **0.876** | **0.783** | **0.946** | **0.965** | **0.997** |

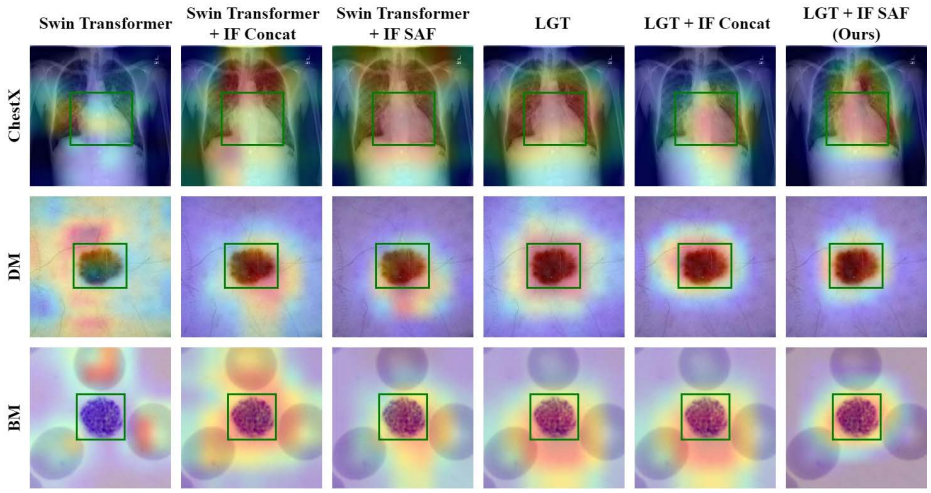


**Fig. 4.** GRAD-CAM visualizations of the six ablation study configurations.

Figure 4 illustrates qualitative examples corresponding to the six rows in Table 3. The addition of cross-layer information enhances the Swin Transformer's ability to understand contextual information (Swin Transformer + IF Concat), thereby allowing the model to focus on the Region of Interest (ROI). Furthermore, integrating the SAF module as a plug-in effectively propagates essential information (Swin Transformer + IF SAF), further refining the model's focus. The LGT module extracts crucial contextual information by focusing on the ROI, although it extends slightly beyond, which the SAF module addresses (LGT + IF SAF). These findings indicate that Med-Former (LGT + IF SAF) effectively captures the essential information necessary for diagnosis.

## 4 Conclusion

We introduced Med-Former, a transformer-based architecture tailored for medical image classification. Addressing limitations in existing models regarding the propagation of essential information from earlier layers and enhancing the feature extraction capability at both local and global levels, we designed the Local-Global Transformer (LGT) and Spatial Attention Fusion (SAF) modules. These modules enable Med-Former to effectively learn both local and global information, facilitating the propagation of essential information through the network. We evaluated our approach on three distinct medical image classification tasks: thoracic disease classification from chest X-rays, skin lesion classification from dermoscopic images, and blood cell classification from microscopic images, respectively. Across these tasks, our approach consistently outperformed other transformer-based architectures and state-of-the-art methods.

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

## References

1. Abdou, M.A.: Literature review: Efficient deep neural networks techniques for medical image analysis. Neural Computing and Applications **34**(8), 5791–5812 (2022)
2. Chan, H.P., Hadjiiski, L.M., Samala, R.K.: Computer-aided diagnosis in the era of deep learning. Medical physics **47**(5), e218–e227 (2020)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Ge, X., Qu, Y., Shang, C., Yang, L., Shen, Q.: A self-adaptive discriminative autoencoder for medical applications. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8875–8886 (2022)
5. Gong, X., Xia, X., Zhu, W., Zhang, B., Doermann, D., Zhuo, L.: Deformable gabor feature networks for biomedical image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4004–4012 (2021)
6. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence **45**(1), 87–110 (2022)
7. 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. pp. 770–778 (2016)
8. 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. pp. 4700–4708 (2017)
9. Kamal, U., Zunaed, M., Nizam, N.B., Hasan, T.: Anatomy-xnet: An anatomy aware convolutional neural network for thoracic disease classification in chest x-rays. IEEE Journal of Biomedical and Health Informatics **26**(11), 5518–5528 (2022)

10. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
11. Lei, Y., Wu, Z., Li, Z., Yang, Y., Liang, Z.: Bp-capsnet: An image-based deep learning method for medical diagnosis. Applied Soft Computing **146**, 110683 (2023)
12. 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/CVF international conference on computer vision. pp. 10012–10022 (2021)
13. Luo, L., Yu, L., Chen, H., Liu, Q., Wang, X., Xu, J., Heng, P.A.: Deep mining external imperfect data for chest x-ray disease screening. IEEE transactions on medical imaging **39**(11), 3583–3594 (2020)
14. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)
15. Nie, W., Zhang, C., Song, D., Bai, Y., Xie, K., Liu, A.A.: Chest x-ray image classification: A causal perspective. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 25–35. Springer (2023)
16. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
18. Wang, S., Liu, X., Zhao, J., Liu, Y., Liu, S., Liu, Y., Zhao, J.: Computer auxiliary diagnosis technique of detecting cholangiocarcinoma based on medical imaging: A review. Computer Methods and Programs in Biomedicine **208**, 106265 (2021)
19. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
20. 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. Scientific Data **10**(1), 41 (2023)