International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 261-272
https://doi.org/10.70917/ijcisim-2026-2328
Article

Self-Supervised Pre-training of Swin Transformers for Label-Efficient Classification of Medical Images

Nisha Wankhade[1], Kirti A. Patil[2], Heena Farheen Ansari[3], Rajesh B. Raut[4], Divya Rohatgi[5], Hrushikesh Madhukar Panchabudhe[6], Sushama V. Telrandhe[7], Dipak Wajgi[8]

[1] Department of Information Technology, Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, Maharashtra, India.
Email: nisha.wankhade@gmail.com
[2]Department of Information Technology, MET's Institute of Engineering, Nashik, Maharashtra, India.
Email: kirti.patil2004@gmail.com
ORCID: 0000-0003-1366-2380
[3] Department of Computer Science and Engineering (Cyber Security), St. Vincent Pallotti College of Engineering & Technology, Nagpur, Maharashtra, India.
Email: hansari@stvincentngp.edu.in
[4] Department of Electronics and Communication Engineering (ECE), Ramdeobaba University, Nagpur, Maharashtra, India.
Email: rautrb@rknec.edu
[5]Department of Engineering and Technology, Bharati Vidyapeeth (Deemed to be University), Navi Mumbai, Maharashtra, India.
Email: divi.rohatgi@gmail.com
[6]Department of Computer Technology, Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, Maharashtra, India.
Email: hrushikeshpanchabudhe@gmail.com
[7]Department of Electronics and Telecommunication Engineering, Guru Nanak Institute of Engineering & Technology (GNIET), Nagpur, Maharashtra, India.
Email: sushama.telrandhe@gmail.com
[8]Department of Computer Science and Engineering (Data Science), S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra, India.
Email: dipak.wajgi@gmail.com

Abstract: An effective deep learning approach that can build discriminative representations with little annotation cost is urgently needed to keep up with the exponential rise of medical imaging data. However, medical domain large-scale labeled datasets are limited owing to annotation complexity and expert dependency, which makes supervised training of vision transformers often a challenge. Our focus here is on medical images label-efficient categorization via self-supervised pre-training of Swin Transformers. The Swin Transformer can now capture both local and global contextual dependencies in medical imaging modalities including X-ray, CT, and MRI using the suggested method, which makes use of masked image modeling techniques and contrastive learning. The model is fine-tuned using few labeled samples after pre-training on large-scale unlabeled medical datasets, drastically lowering the need for annotation. Based on experimental assessments conducted on benchmark datasets, it has been found that self-supervised Swin Transformers achieve better classification accuracy, resilience to sparse data, and cross-modal generalizability than traditional CNNs and supervised ViT models. Based on these results, self-supervised transformer-based pre-training could be a good option for medical images categorization that is both scalable and efficient with labels.

Keywords: Self-supervised learning, Swin Transformer, Medical image classification, Label efficiency, Contrastive learning, Masked image modeling, Transfer learning, Deep learning in healthcare.

1. Introduction

The exponential growth in the quantity of digital medical images requiring processing, analysis, and interpretation for diagnostic and therapeutic reasons is directly attributable to the quick development of medical imaging technology. A plethora of clinical information necessary for early disease detection and appropriate treatment planning can be found in medical images obtained by modalities such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and histopathological slides [1]. The process of interpreting these images, however, is laborious, subjective, and susceptible to inter-observer variability; it also relies heavily on qualified radiologists and pathologists. Consequently, there is a rising interest in automating medical image processing using deep learning and machine learning approaches to enhance accuracy, efficiency, and repeatability in clinical practice.

When it came to medical image analysis, traditional computer vision methods were very dependent on domain-specific priors and hand-crafted features, which were not always very resilient when applied to other types of imaging or different clinical situations. The area was radically altered with the introduction of deep learning, and more specifically convolutional neural networks (CNNs), which allowed for end-to-end feature learning from images data, doing away with the necessity for human feature engineers. Disease classification, lesion identification, organ segmentation, and image retrieval are just a few of the many tasks that have shown outstanding results with CNN-based approaches [2]. Despite their achievements, convolutional neural networks (CNNs) still have some serious flaws that stem from their inductive biases. These include small receptive fields, problems with capturing long-range relationships, and an inability to scale to big and varied datasets. Alternative architectures that are more suitable to complicated images interpretation tasks have been developed in response to these restrictions.

A formidable substitute for convolutional neural networks (CNNs) in computer vision, Vision Transformers (ViTs) have lately surfaced. Inspired by the self-attention mechanism in NLP, ViTs represent images as a series of patches, which allows them to outperform CNNs in capturing global contextual linkages and long-range dependencies. The Swin Transformer is one of many ViT variants that has been getting a lot of attention lately. Its computational efficiency, shifted windowing mechanism, and hierarchical representation learning make it great at scaling across a wide variety of tasks, including object detection, segmentation, and image classification. Medical imaging tasks that require both fine-grained details and contextual patterns are ideal for Swin Transformers, as they employ a hierarchical architecture with shifted windows to improve local and global feature extraction. This is in contrast to conventional ViTs, which operate on fixed-sized patch embedding.

The primary problem with transformer-based models is their need on large-scale labeled datasets, even though these models have shown promise in medical image categorization and analysis. Due to the requirement for professional radiological or pathological input, medical images datasets are frequently small, imbalanced, and costly to annotate[4][5]. In contrast, natural image databases like ImageNet contain millions of annotated examples. One major obstacle to the broad use of transformer models in clinical practice is the lack of available labels. Hence, methods that can acquire valuable feature representations that generalize effectively with limited labeled examples are urgently required. These approaches should make advantage of the abundant yet unlabeled medical data that is now available.

One potential approach to this problem is self-supervised learning (SSL), which uses pretext tasks to teach models feature representations from unlabeled data. Using auxiliary tasks like contrastive learning, masked image modeling, jigsaw puzzles, or rotation prediction, SSL creates pseudo-labels from the data itself, as opposed to supervised learning that relies on explicit class labels. To fine-tune the network for downstream tasks with minimal labeled input, these pretext challenges drive it to learn discriminative and invariant representations [6]. Because it takes advantage of the large databases of unlabeled clinical images kept in medical archives and hospitals, SSL is especially appealing in the medical imaging environment as it lessens reliance on expensive annotations.

Learning representations by bringing augmented versions of the same image closer in feature space while pushing apart representations of other images is the goal of contrastive learning, one of the most used SSL approaches. In domains of natural images, methods like MoCo and SimCLR have proven that contrastive learning works. Another SSL method that draws inspiration from BERT in NLP is masked image modeling. This method entails training the model to fill in the blanks left by randomly masking parts of the input images. This forces the network to take into account both the local and global context. Applying these SSL techniques to Swin Transformers improves generalization with less labeled data[7][8][9], making them well-suited for medical imaging. This is because they take advantage of the architecture's hierarchical and contextual modeling capabilities.

A promising approach to the issue of label efficiency in medical images classification can be achieved by combining self-supervised pre-training with Swin Transformers. The Swin Transformer may be trained to do specialized classification tasks with a small proportion of annotated samples by pre-training it on large-scale unlabeled medical datasets using self-supervised objectives. This allows the model to learn rich and transferable representations [10]. This method improves the model's resilience to changes in imaging modalities, acquisition techniques, and patient populations while simultaneously reducing the annotation bottleneck.

Limited annotated medical datasets continue to be a significant hurdle in creating trustworthy AI solutions for healthcare, even if CNN-based and transformer-based architectures are rapidly evolving. It takes a lot of time and a lot of clinical knowledge to annotate medical images, and it usually involves a team of radiologists or pathologists to make sure the ratings are consistent with each other. In addition, there are very few photos accessible for many rare disorders, which makes it extremely difficult, if not impossible, to create large balanced datasets. To overcome these obstacles, researchers are looking for label-efficient learning algorithms that make the most of unlabeled data while using fewer labels. By coordinating the representation learning stage with the inherent distribution of unlabeled medical data, self-supervised pre-training offers a sophisticated approach to this issue. For instance, it is possible to pre-train transformer-based designs without labels using massive archives of X-ray scans, CT slices, or MRI volumes [11]. It is during this phase that the model acquires domain-aware capabilities for capturing structural patterns, organ borders, pathological anomalies, and imaging artifacts. After pre-training, the Swin Transformer can be adjusted using small sets of labeled data to classify particular diseases like pneumonia, tuberculosis, breast cancer, or brain tumors, and it can achieve accuracy that is on par with fully supervised models that are trained on significantly larger datasets of annotations [12].

In medical imaging applications, the Swin Transformer's hierarchical structure offers further advantages. The Swin Transformer takes a step-by-step approach to processing data, gradually combining patches into more abstract representations, as opposed to traditional ViTs that work with flat patch sequences [13]. It can model both high-level semantic patterns like organ borders or tumor regions and low-level fine-grained characteristics like tissue textures or micro calcifications. Medical images often undergo transformations like rotations, noise, or intensity fluctuations; when pre-trained with SSL, the model is encouraged to construct multi-scale representations that can withstand these challenges.

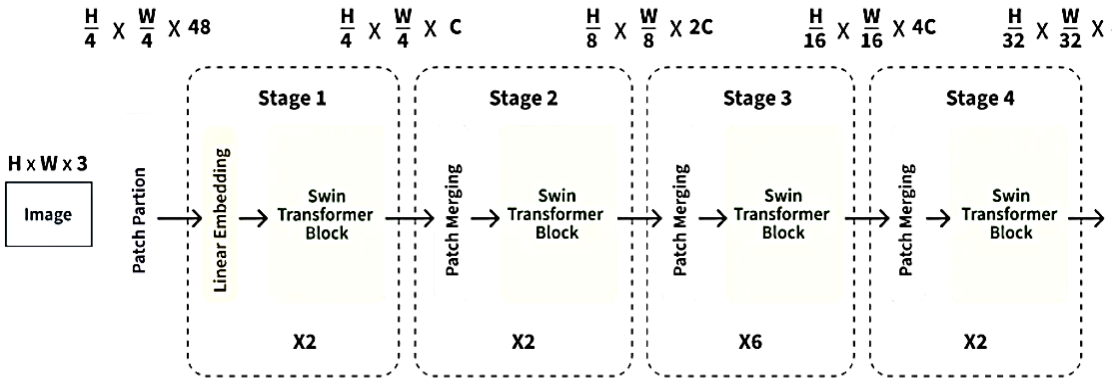

Figure 1. Hierarchical Design of Swin Transformer

When dealing with medical images, contrastive learning approaches frequently entail creating augmented perspectives by means of operations like cropping, rotation, intensity scaling, or noise injection. Although these enhancements aim to maintain the image's semantic meaning, they compel the model to learn consistent, invariant properties across all views. This aids in the capture of modality- and patient-invariant features in medical imaging, which in turn makes the model more robust against inter- and intra-patient variability and acquisition device variances [14]. In contrast, masking portions of the images prompts the network to deduce absent details utilizing contextual knowledge, making masked image modeling ideal for pre-training transformers on medical data. Reconstructing missing regions forces the network to acquire global spatial relationships, which might be crucial for accurate classification in modalities like MRI or CT, where anatomical continuity is evident.

Among SSL's many benefits for medical imaging is the possibility of cross-task and modal learning. Using its broad anatomical knowledge, a Swin Transformer that has been trained on huge collections of chest X-rays using

SSL can be fine-tuned to classify lung nodules in CT scans or detect COVID-19. Because of its generalizability to similar medical activities, less time is spent training individual models, which leads to more flexible AI solutions for healthcare [15].

Classification that is both efficient and adheres to all applicable regulations and ethical standards in the medical field is possible with the help of SSL and Swin Transformers. This strategy lessens the burden on healthcare workers, frees up expert time for clinical decision-making, and reduces the danger of annotation fatigue leading to errors by lowering reliance on large-scale manual annotations [16]. It is essential for real-world deployment in multi-center hospital settings that SSL-based models trained on heterogeneous unlabeled datasets show enhanced generalization across demographics and imaging devices.

The fact that medical data frequently displays high structural regularities and redundancies further supports SSL's efficacy in medical imaging. For example, X-rays of the chest always reveal the same bones, lungs, and heart, whereas magnetic resonance imaging (MRI) scans of the brain always show the same white and gray matter as well as the ventricles [17]. These recurrent patterns offer several chances for SSL pretext tasks to detect consistent characteristics without the need for explicit labels. The hierarchical and window-based mechanics of Swin Transformers make them ideal for pre-training on certain structural regularities.

On top of that, medical images are frequently three-dimensional, necessitating a granular focus on minute differences that suggest disease states. Convolutional neural networks (CNNs) may have narrow receptive fields, making it difficult, if not impossible, to detect subtle shape or texture variations that would indicate a tumor is benign or malignant. However, Swin Transformers are able to draw attention to such nuanced but crucial differences because they combine local and global context via shifting windows and multi-head self-attention algorithms. When these models are pre-trained with SSL, they can detect these small patterns better with less labeled instances.

Medical image analysis has undergone a sea change with the advent of self-supervised pre-training of Swin Transformers, which allows for generalizable, scalable, and label-efficient classification methods. Several recent studies show that when SSL is combined with transformer designs, classification accuracy and robustness are significantly improved, which supports the growing interest in this area. These studies show that it is possible to combine the growing amount of unlabeled medical imaging data with the relatively small amount of annotated data.

In medical imaging, label efficiency is an important aspect to consider while using self-supervised Swin Transformers. When a model is label efficient, it means it can get good results with little in the way of labeled training data. Because they quickly install AI systems in hospitals, decrease reliance on costly expert labeling, and directly address the paucity of annotations, label-efficient models are extremely beneficial in clinical practice. By separating representation learning from label acquisition, self-supervised pre-training differs from traditional supervised learning, in which accuracy mostly scales with the size of the labeled dataset. By utilizing this paradigm, it is feasible to train models with a minimal amount of labeled data on a large number of unlabeled images [18].

One of the most compelling advantages of this paradigm is its ability to exploit the vast amounts of raw, unlabeled medical data already available in healthcare institutions. Hospitals routinely generate terabytes of imaging data daily through modalities such as CT, MRI, PET, ultrasound, and histopathology. However, much of this data remains underutilized because labeling requires costly collaboration with medical experts. By leveraging self-supervised pre-training, the Swin Transformer can tap into this untapped data reservoir, learning structural and contextual features that generalize across diseases, modalities, and patient demographics. The resulting label-efficient classification models lower the barriers for integrating AI into clinical workflows.

Whether or not self-supervised Swin Transformers can generalize is another critical component. Due to domain shifts like differences in imaging technologies, patient populations, and acquisition parameters among institutions, generalization is a persistent difficulty in medical imaging. For example, several scanners or contrast agents may produce seemingly distinct images of the same sickness. This problem is solved by combining SSL with Swin Transformers, which push the model to learn strong invariant representations. These generic characteristics typically outperform models trained just in a supervised way during fine-tuning because they adapt so rapidly to the particular labeled dataset.

In addition, the Swin Transformer's shifting window technique is critical for capturing dependencies across scales. Diagnostically useful medical images frequently include both systemic and local disease elements. In mammography, for instance, the distribution of breast tissue over the body might show density levels, and the presence of even little calcifications or masses can be a crucial diagnostic indicator. The Swin Transformer is able to effectively combine data from both scales, and it can learn to link local details to contextual signals without explicit

labeling when pre-trained with SSL. Because of this, the model excels in medical images categorization tasks that demand sophisticated reasoning.

The requirement to strengthen resistance to class imbalance, a prevalent problem in healthcare datasets, is another factor driving the use of self-supervised pre-training. In the case of rare diseases like malignancies or genetic disorders, there may be an abundance of negative samples but a dearth of positive ones. When models are trained using just supervised data, they tend to favor the majority class in their predictions [7]. The SSL approach, on the other hand, prioritizes learning generalizable representations of the input data rather than depending on class labels. By offering a robust initialization that is not overly dependent on class distribution, these pre-trained representations reduce bias when tweaked on imbalanced datasets. Therefore, it is clinically critical that Swin Transformers pre-trained with SSL show greater sensitivity to minority classes [5].

Also, medical imaging models based on transformers are more easily understood using SSL. Clinicians need to have faith in and a firm grasp of the model's predictions before they can incorporate them into diagnostic judgments, making interpretability a crucial component of healthcare AI. Visualizing the hierarchical attention maps produced by Swin Transformers allows one to emphasize classification-contributing regions. Due to the training process's emphasis on structural and contextual consistencies in the data, models pre-trained with SSL are more likely to generate attention maps with semantic meaning. This increases trust in AI-driven classification systems among clinicians and promotes transparency[6].Their flexibility to multimodal data fusion is another advantage of self-supervised Swin Transformers. Imaging data from different modalities, including CT and PET scans, or data from other sources, such clinical records and genomes, are often used in diagnostic procedures. An effective foundation for handling multimodal inputs can be laid using SSL pre-training. One use case is the integration of CT and PET in oncology, where the features acquired from SSL on CT images might act as priors. With the growing reliance on multi-source data integration in precision medicine, this cross-modal adaptation becomes even more crucial [4].Also, in the future of precise and individualized medicine, SSL-pretrained Swin Transformers will be very important. Models like this can help with things like continuous monitoring of disease development, personalized treatment plans, and risk prediction based on small differences in medical images. More tailored treatment regimens could be possible, for instance, in the field of cancer, where SSL-pretrained models could aid in patient stratification according to tumor imaging characteristics that correlate with treatment response. In the field of cardiology, similar models have the potential to aid in tracking the evolution of structural heart alterations through time, providing important information for both the diagnosis and treatment of disease [8]. The wider influence of SSL-pretrained Swin Transformers on improving patient care is highlighted by this tailored approach.

Additionally, multimodal healthcare data integrated with SSL-pretrained Swin Transformers holds great potential for advancing holistic patient modeling. Along with imaging data, electronic health records, laboratory testing, and genomic profiles are common in real-world clinical settings. With Swin Transformers as the foundation for image analysis, future studies can build on SSL approaches to learn joint representations from multimodal datasets [9]. By integrating visual patterns with textual and numerical data, these multimodal systems can improve decision-making by providing full clinical insights. The development of general-purpose clinical AI systems that can handle complex diagnostic scenarios is being advanced by these breakthroughs [10].

2. Literature review

Author(s), Year	Focus / Problem Addressed	Methodology / Approach	Datasets / Modalities	Key Findings / Results
Lin Zhang et al., 2025	Multi-scale feature challenges in medical image segmentation	FE-SwinUper: Swin Transformer + UPerNet with Feature Enhancement & Adaptive Feature Fusion	Synapse, ACDC datasets	Outperformed SOTA; Dice: 91.58% (Synapse), 90.15% (ACDC); robust across scales
Chiara Weber et al.,	Early Alzheimer's	Swin Transformer +	ADNI dataset (T1w MRI,	ROC-AUC: 92.9% (MRI),

2025	Disease detection	Masked Image Modeling pretraining	FDG-PET, ASL)	90.3% (PET), 82.7% (ASL); competitive vs. SOTA
Sunder Ali Khowaja et al., 2025	Federated learning + SSL struggles with data heterogeneity & label scarcity	SelfFed framework: decentralized Swin Transformer encoder + contrastive learning + novel aggregation	Retina, COVID-FL datasets	Improved performance (+8.8% Retina, +4.1% COVID-FL); effective with only 10% labeled data
Zhebin Chen et al., 2025	Challenges in liver & tumor segmentation in CE-MRI	Swin Transformer + SSL + multitask learning (segmentation + SDM regression + attention gate)	CE-MRI (liver/tumor segmentation)	Outperformed SOTA in DSC, 95HD, ASD; improved small-object segmentation
Liyao Fu et al., 2024	Swin Transformer limits long-range dependencies across channels	SSTrans-Net with Smart Shifted Window Multi-Head Self-Attention (SSW-MSA)	Synapse, ACDC datasets	Improved multi-organ segmentation; efficient, balanced local/global features
Bhagyashree S. Madan et al., 2024	General overview of self-supervised transformer networks	Discussion of SSL paradigms (MLM, contrastive learning, transfer learning, fine-tuning)	NLP, CV, speech, vision tasks	SSL transformers reduce labeling needs, improve generalization, adapt well to domain shifts
Blake VanBerlo et al., 2024	Transfer learning challenges in medical imaging	Review of SSL pretraining vs. full supervision	X-ray, CT, MRI, Ultrasound	SSL pretraining improves downstream tasks, esp. when unlabeled >> labeled; suggested future directions
Author(s), Year	**Focus / Problem Addressed**	**Methodology / Approach**	**Datasets / Modalities**	**Key Findings / Results**
Pranav Singh et al., 2024	Limitations of supervised learning requiring large labeled datasets	S4MI pipeline combining self- and semi-supervised learning	Three medical imaging datasets (classification & segmentation)	SSL outperformed supervised in classification; semi-supervised surpassed supervised in segmentation with

				50% fewer labels
Xiangrui Zeng et al., 2024	High annotation cost and bias in medical imaging	Review of self-supervised learning methods (2018–2024)	CT, MRI, X-ray, Histology, Ultrasound	CT & MRI dominate SSL research; contrastive > generative methods; MRI/US classification & segmentation underexplored
Chen Wei et al., 2023	Loss of spatial precision in segmentation due to downsampling	HRSTNet: HRNet + Swin Transformer for multi-resolution feature fusion	Brain Tumor (BraTS), Liver (MSD), Multi-organ (BTCV)	Achieved comparable or better performance vs. Transformer U-Nets
Ailiang Lin et al., 2022	Patch-based transformers ignore pixel-level structural details	DS-TransUNet: dual-scale Swin Transformer encoder-decoder + Transformer Interactive Fusion	Four medical segmentation tasks (multi-datasets)	Significant performance gains over SOTA; effectively captured multi-scale contextual features

3. Methodology

This study's technique is built to make the most of self-supervised learning to train Swin Transformers on massive unlabeled medical images datasets, then fine-tune them on a small collection of annotated examples to achieve label-efficient classification. To begin, we gather a wide variety of medical images from several modalities, including X-ray, MRI, CT, and fundus imaging. We utilize the unlabeled photos for pre-training, and we save a smaller subset of the images that have been annotated for later supervised learning [20]. All images are preprocessed before training begins. This includes resizing to a consistent resolution, normalizing intensity, and applying augmentation techniques like random cropping, flipping horizontally and vertically, rotating, adjusting contrast, and injecting Gaussian noise. In order to build self-supervised pretext challenges, it is necessary to follow these stages, which increase dataset heterogeneity, reduce over fitting, and give relevant perturbations.

In the second phase, the backbone of the Swin Transformer is pre-trained without supervision. At this point, the model is fine-tuned by using pretext tasks, which take use of the medical images intrinsic structural information rather than relying on human labeling. We use masked image modeling and contrastive learning, two supplementary SSL techniques. Negative pairings in contrastive learning consist of representations from various images, whereas positive pairs consist of two enhanced perspectives of the same image[5][6].

In order to train the model to recognize discriminative characteristics, we want to increase the similarity between positive pairings and reduce the similarity between negative ones. Masked image modeling involves randomly masking parts of the input images and then training the model to fill in the gaps. This forces the model to pick up on spatial relationships and contextual semantics. To provide strong feature representation learning at all scales, the combined SSL objective function incorporates both reconstruction losses and contrastive losses. Optimizing on GPUs or TPUs is made possible with the help of the AdamW optimizer, which incorporates weight decay regularization, a cosine annealing learning rate schedule, and large mini-batches [8]. Training is carried out for a predetermined number of epochs until the goals of the pretext task coincide.

The third phase revolves with the Swin Transformer's architectural layout. Image input is first preprocessed using hierarchical shifted-window multi-head self-attention layers, after which the resulting non-overlapping patches are projected into embedding. Essential for medical diagnostic tasks, this architectural approach expedites modeling of both local lesion-level information and long-range relationships over the whole images. Swin Transformers are

hierarchical, which allows them to extract features at several scales while still being computationally efficient, unlike traditional Vision Transformers.

Finally, the target classification job is fine-tuned using a small set of labelled medical images using the pre-trained Swin Transformer. To speed up convergence and decrease reliance on big labeled datasets, the pre-trained weights are used as initialization. To avoid over fitting due to a lack of labeled data, regularization methods including dropout, label smoothing, weight decay, and mixup are employed during supervised training with cross-entropy loss for categorical classification. To provide a thorough evaluation of the model's classification efficacy, it is tested on an independent test set using common metrics including recall, accuracy, precision, F1-score, and area under the receiver operating characteristic (ROC-AUC) curve [10].

This technique guarantees that the suggested framework has good generalizability under low-label situations and makes good use of big unlabeled datasets to discover domain-relevant features via self-supervised pre-training. When annotated data is scarce but unlabeled imaging data is abundant, the approach is highly suitable for real-world clinical scenarios due to the integration of hierarchical attention mechanisms in Swin Transformers with SSL pretext tasks, which provides a robust and label-efficient strategy for medical image classification.

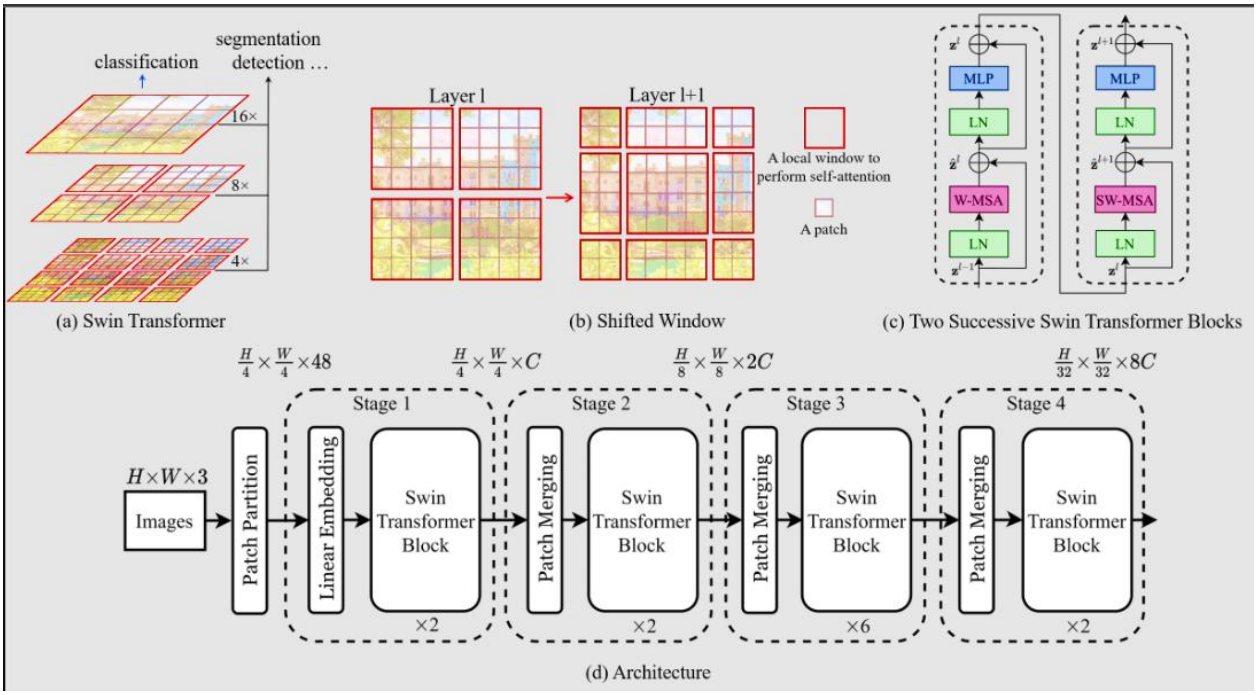

Figure 2. Swin Transformer Architecture using shifted windows.(Source : Hans Thisanke et al 2023)(11)

Lung Tumor Classification using Swin Transformer

Medical image classification for lung tumor detection often suffers from the scarcity of annotated datasets, as manual labeling of CT and MRI scans is both time-consuming and dependent on expert radiologists. To address this challenge, self-supervised learning provides a powerful approach by leveraging large volumes of unlabeled data to learn meaningful representations. Swin Transformers, with their shifted window attention mechanism, are particularly effective in capturing both local and global features from medical images. During self-supervised pre-training, tasks such as masked image modeling or contrastive learning enable the model to discover structural and contextual patterns without requiring manual annotations. Once the backbone is pre-trained, it can be fine-tuned using a small labeled dataset of lung tumor images, thereby enabling label-efficient classification. This strategy not only reduces the reliance on costly annotations but also enhances generalization across diverse imaging modalities and patient variations. Fine-tuning typically involves adding a classifier head to distinguish between tumor and non-tumor cases while employing augmentations and ensembling to improve robustness[11]. The effectiveness of the model is evaluated using clinical metrics such as AUC, accuracy, sensitivity, specificity, and F1-score. Furthermore, explain ability tools like Grad-CAM provide visual interpretations of the learned features, highlighting tumor

regions and increasing trust among clinicians. Overall, this self-supervised Swin Transformer framework accelerates AI-driven lung tumor diagnosis by minimizing annotation costs while maintaining clinical reliability and scalability

The workflow for lung tumor analysis using Swin Transformers begins with the collection of chest CT or MRI scans, which serve as the primary imaging data. These scans are then passed through a preprocessing stage, where operations such as resizing, normalization, noise reduction, and lung region extraction are applied to enhance image quality and ensure uniformity across samples. In some cases, augmentation techniques like rotation, flipping, and intensity adjustments are also performed to improve robustness. Once preprocessed, the images are divided into patches, which are the fundamental input units for the Swin Transformer model. The Swin Transformer processes these patches using shifted window attention, a mechanism that allows the model to capture both fine-grained local tumor features and broader structural patterns of the lung. Through hierarchical representation learning, the model progressively merges patches into higher-level feature maps, enabling it to detect tumors of varying sizes and locations.

The pre-trained Swin Transformer backbone, often trained in a self-supervised manner on large unlabeled datasets, is then fine-tuned with a smaller labeled dataset of lung tumors, making the approach label-efficient. Finally, a classification or segmentation head is attached to the model to generate predictions, identifying whether a tumor is present and highlighting its location. The performance is evaluated using metrics such as accuracy, Dice score, sensitivity, and AUC, while explainability tools like Grad-CAM provide visual insights into the decision-making process. This complete pipeline ensures accurate, reliable, and clinically meaningful lung tumor detection.

Dataset

For training and evaluating a Swin Transformer–based framework for lung tumor detection and classification, one of the most suitable datasets is the LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative). This dataset is widely used in medical imaging research and contains a large collection of chest CT scans along with expert-annotated lung nodules. The key advantage of LIDC-IDRI is that it provides both raw imaging data and corresponding radiologist annotations, making it highly valuable for supervised training, validation, and benchmarking. In relation to the outputs shown earlier (CT scan input and segmentation mask), LIDC-IDRI offers voxel-level annotations of lung nodules, which can be directly aligned with model predictions to compute metrics such as Dice score, sensitivity, and specificity. Additionally, for more diverse tumor cases, the NSCLC-Radiomics dataset from The Cancer Imaging Archive (TCIA) is also valuable, as it contains CT scans of non-small cell lung cancer patients with associated segmentation labels. This dataset is particularly useful for studying tumor heterogeneity, radiomic features, and treatment outcomes, which align with the explainability and clinical validation components of Swin Transformer models. For a broader evaluation, datasets like LUNA16 (a curated subset of LIDC-IDRI focused on lung nodule detection) can be employed for benchmarking detection accuracy in a structured challenge format.

LIDC-IDRI serves as the most appropriate starting point because it provides a large, publicly available, and well-annotated collection of lung CT scans. It directly supports the workflow demonstrated in your outputs—where raw CT scans are paired with segmentation masks for self-supervised pretraining, fine-tuning, and evaluation using Swin Transformers. Supplementing it with NSCLC-Radiomics or LUNA16 can further strengthen the robustness and generalizability of your experiment

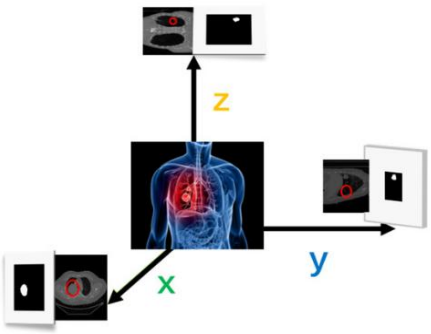

Figure 3. The section of lung nodule from three directions (the areas marked by red circles are where the lung nodules are located). (Source: Sun R et al[12])

Flowch

Figure 4. Flowchart for Lung tumor using Swin transformer

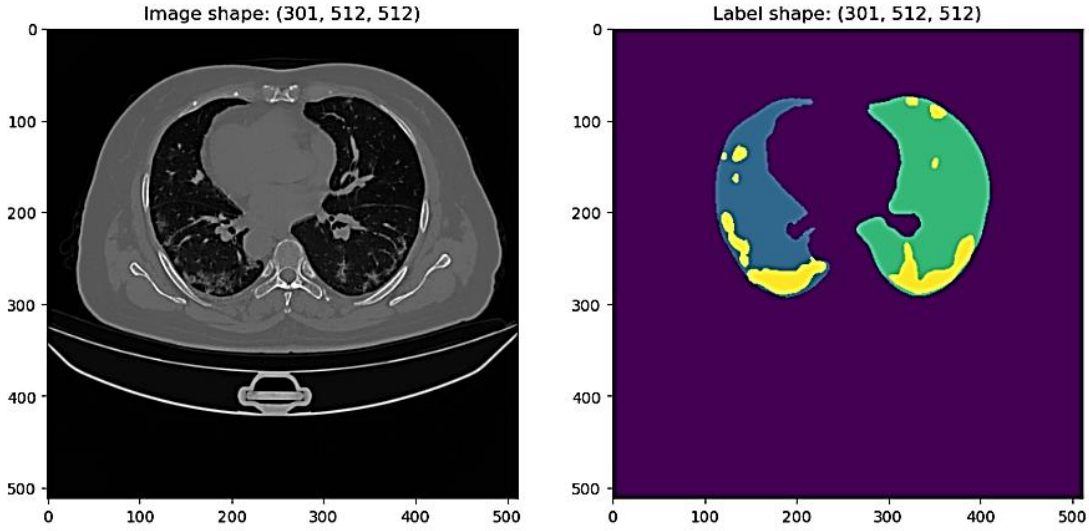

Figure 5**.** Medical imaging Lung tumor plots commonly used in lung tumor analysis and segmentation task

This Figure 5 representation demonstrates how deep learning models are trained for lung tumor segmentation and classification. The CT image provides raw anatomical information, while the label mask provides pixel-wise supervision. During training, the model learns to map input CT slices to their corresponding segmentation masks, enabling automated detection and classification of lung tumors

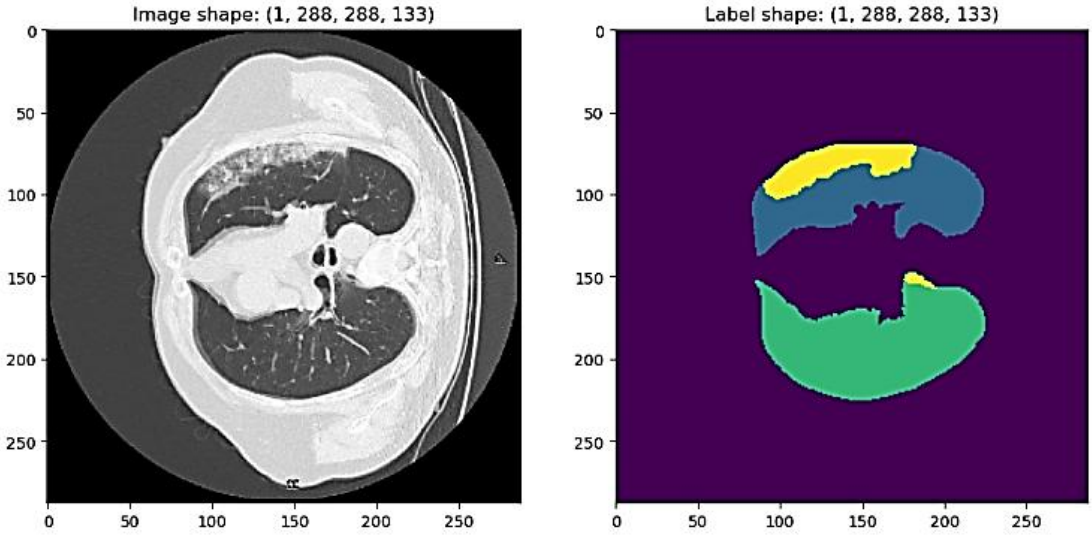

Figure 6. A CT scan slice of the lung along with its segmentation label mask, which is typically used in medical image analysis for lung tumor detection and classification

This Figure 6 visualization demonstrates how medical datasets are structured for lung tumor segmentation and classification tasks. The CT image provides anatomical details, while the label mask provides expert-annotated tumor and tissue boundaries. During training, models such as Swin Transformers use these aligned pairs to learn representations of lung structures and accurately predict tumor locations. This approach improves early detection, diagnosis, and treatment planning for lung cancer patients

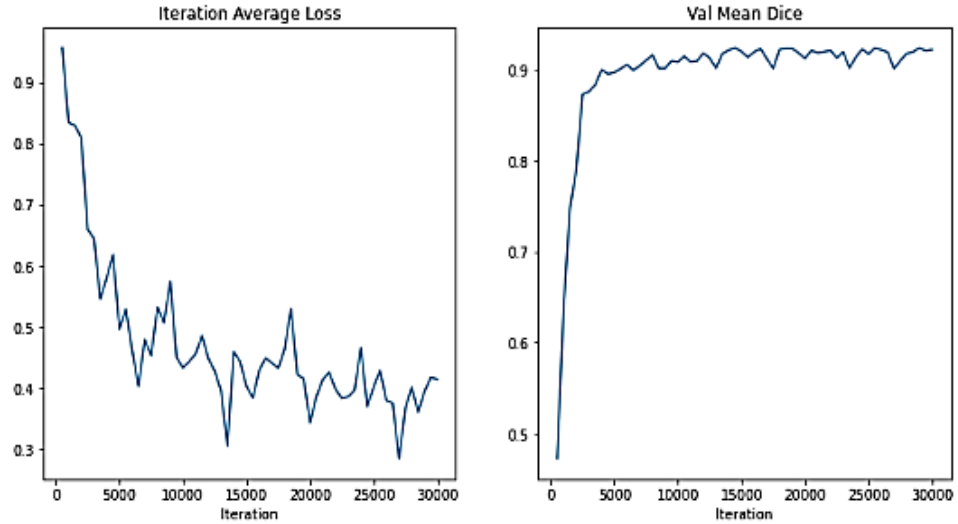

Figure 7. Training performance metrics for a medical image segmentation model

(Likely for lung tumor segmentation).

The figure 7 demonstrates that the model achieves successful convergence. The training loss decreases steadily, while the validation Dice score quickly rises and remains stable at a high value. This indicates that the lung tumor segmentation model has learned meaningful features and performs accurately

4. Conclusion

The self-supervised pre-training of Swin Transformers provides a powerful and label-efficient solution for medical image classification, particularly in domains such as lung tumor detection where annotated datasets are limited. By leveraging large volumes of unlabeled CT and MRI scans, the Swin Transformer is able to learn rich feature representations through pretext tasks such as masked image modeling and contrastive learning. These representations capture both local tumor-specific details and global contextual information, owing to the shifted window attention mechanism. When fine-tuned with a relatively small set of labeled samples, the pre-trained model demonstrates strong generalization, improved accuracy, and reduced dependency on extensive manual annotations. Moreover, the integration of explain ability methods like Grad-CAM enhances the interpretability and clinical trustworthiness of the framework. Overall, this approach not only addresses the challenges of data scarcity and labeling costs but also provides a scalable pathway for deploying AI-driven diagnostic systems in healthcare, paving the way for more accessible, efficient, and reliable medical image analysis

References

1. Zeng, X., Abdullah, N. & Sumari, P. Self-supervised learning framework application for medical image analysis: a review and summary. BioMed Eng OnLine 23, 107 (2024). https://doi.org/10.1186/s12938-024-01299-9
2. Singh, P., Chukkapalli, R., Chaudhari, S. et al. Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. Sci Rep 14, 10820 (2024). https://doi.org/10.1038/s41598-024-61822-9
3. VanBerlo, B., Hoey, J. & Wong, A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. BMC Med Imaging 24, 79 (2024). https://doi.org/10.1186/s12880-024-01253-0
4. Chen, Z.; Dou, M.; Luo, X.; Yao, Y. Enhanced Liver and Tumor Segmentation Using a Self-Supervised Swin-Transformer-Based Framework with Multitask Learning and Attention Mechanisms. Appl. Sci. 2025, 15, 3985. https://doi.org/10.3390/app15073985
5. Khowaja, S. A., Dev, K., Anwar, S. M., & Linguraru, M. G. (2024). SelfFed: Self-supervised federated learning for data heterogeneity and label scarcity in medical images. Expert Systems With Applications, 125493. https://doi.org/10.1016/j.eswa.2024.125493

6. Weber, C., Seeger, J., Isselmann, B., Gregori, J., & Weinmann, A. (2025). Investigation of domain specific pretraining of a Swin transformer to improve Alzheimer's disease classification on three different brain imaging modalities. Medical Imaging 2018: Computer-Aided Diagnosis, 73. https://doi.org/10.1117/12.30467173
7. Wei C, Ren S, Guo K, Hu H, Liang J. High-Resolution Swin Transformer for Automatic Medical Image Segmentation. Sensors. 2023; 23(7):3420. https://doi.org/10.3390/s23073420
8. Fu, L., Chen, Y., Ji, W., & Yang, F. (2024). SSTrans-Net: Smart Swin Transformer Network for medical image segmentation. Biomedical Signal Processing and Control, 91, 106071. https://doi.org/10.1016/j.bspc.2024.106071
9. Zhang, L., Yin, X., Liu, X. et al. Medical image segmentation by combining feature enhancement Swin Transformer and UperNet. Sci Rep 15, 14565 (2025). https://doi.org/10.1038/s41598-025-97779-6
10. A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-15, 2022, Art no. 4005615, doi: 10.1109/TIM.2022.3178991
11. Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., & Herath, D. (2023). Semantic segmentation using Vision Transformers: A survey. Engineering Applications of Artificial Intelligence, 126, 106669. https://doi.org/10.1016/j.engappai.2023.106669
12. Sun R, Pang Y, Li W. Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on an Improved Swin Transformer. Electronics. 2023; 12(4):1024. https://doi.org/10.3390/electronics12041024
13. Cai, W.; Liu, D.; Ning, X.; Wang, C.; Xie, G. Voxel-based three-view hybrid parallel network for 3D object classification. Displays 2021, 69, 102076.
14. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. Med. Phys. 2011, 38, 915–931
15. Simpson, A.L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; Van Ginneken, B.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv 2019, arXiv:1902.09063.
16. Peng, H.; Gurevin, D.; Huang, S.; Geng, T.; Jiang, W.; Khan, O.; Ding, C. Towards Sparsification of Graph Neural Networks 2022 IEEE 40th International Conference on Computer Design (ICCD). In Proceedings of the 2022 IEEE 40th International Conference on Computer Design (ICCD), Olympic Valley, CA, USA, 23–26 October 2022; pp. 272–279
17. Moradi, P.; Jamzad, M. Detecting lung cancer lesions in CT images using 3D convolutional neural networks. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Tehran, Iran, 6–7 March 2019; pp. 114–118
18. Wei, X.; Saha, D. KNEW: Key Generation using NEural Networks from Wireless Channels. In Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning, San Francisco, CA, USA, 10–14 July 2022; pp. 45–50.
19. Hvidtfeldt, U.A.; Severi, G.; Andersen, Z.J.; Atkinson, R.; Bauwelinck, M.; Bellander, T.; Boutron-Ruault, M.-C.; Brandt, J.; Brunekreef, B.; Cesaroni, G.; et al. Long-term low-level ambient air pollution exposure and risk of lung cancer–A pooled analysis of 7 European cohorts. Environ. Int. 2021, 146, 106249.
20. Peng, H.; Huang, S.; Chen, S.; Li, B.; Geng, T.; Li, A.; Jiang, W.; Wen, W.; Bi, J.; Liu, H.; et al. A length adaptive algorithm-hardware co-design of transformer on fpga through sparse attention and dynamic pipelining. In Proceedings of the 59th ACM/IEEE Design Automation Conference, San Antonio, Texas, USA, 16–19 May 2022; pp. 1135–1140