BIRADS Breast Cancer Classification: Swin Transformer Approach on VinDr Mammo Data

Krithi Shailya EE20B038
Department of Electrical Engineering
Indian Institute of Technology Tirupati

Sanket Wathore EE20B041
Department of Electrical Engineering
Indian Institute of Technology Tirupati

***Abstract*—Breast cancer remains a significant global health challenge, demanding accurate screening and diagnosis for effective treatment. This paper presents a novel approach to breast cancer classification utilizing the Swin Transformer architecture on the VinDr Mammo dataset. Leveraging deep learning and computer-assisted detection techniques, our method aims to enhance breast cancer detection and diagnosis. We integrate the Swin Transformer into VinDr for superior performance, overcoming challenges such as data loss and format conversions. Our future scope includes exploring different models, integrating preprocessing techniques, employing ensemble learning, and rigorously evaluating VinDr's performance to enhance its diagnostic capabilities. Overall, our objective is to contribute to improved breast cancer screening and diagnosis, ultimately benefiting patient outcomes and healthcare practices.**

***Index Terms*—VinDr Mammo, Swin Transformer**

I. Introduction

Breast cancer poses a significant global health challenge, representing one of the most prevalent cancers and a leading cause of cancer-related mortality. Early detection is crucial for effective treatment and improved survival rates. Mammography, a widely adopted imaging modality, plays a central role in breast cancer screening programs due to its ability to detect abnormalities in breast tissue. However, the interpretation of mammography images, particularly in identifying potentially malignant lesions, remains a complex and challenging task for radiologists.

In recent years, there has been a notable surge in research focused on leveraging advancements in deep learning and computer-assisted detection and diagnosis (CADe/x) tools to enhance the accuracy and efficiency of mammography interpretation. These AI-based systems have shown promising results in aiding radiologists by providing additional support in detecting and characterizing breast lesions, thereby potentially improving screening outcomes.

However, the development and evaluation of CADe/x algorithms require access to diverse and well-annotated datasets. These datasets serve as the foundation for training, testing, and validating AI models, allowing researchers to assess the performance and generalization capabilities of their algorithms across different populations and clinical settings. Moreover, comprehensive annotations, including detailed classifications such as those provided by the Breast Imaging Reporting and Data System (BI-RADS), are essential for the development of robust and clinically relevant CAD systems.

Recognizing the importance of such resources, a recent endeavor has introduced the VinDr-Mammo dataset, a novel addition to the landscape of publicly available mammography datasets. This dataset, originating from Vietnam, offers a unique contribution by enhancing the diversity of available data and providing extensive annotations at both the breast-level and lesion-level.

In this paper, we provide a detailed overview of the VinDr-Mammo dataset, focusing particularly on its BI-RADS classification scheme, dataset creation process, validation procedures, potential applications, and associated limitations. By shedding light on the capabilities and characteristics of this dataset, we aim to facilitate its effective utilization in advancing research and development efforts aimed at improving breast cancer screening and diagnosis through AI-driven approaches.

Our code can be found at: https://github.com/iikrithii/DLHC_BIRADS_Swin/

II. Background and Literature Review

A. VinDr Mammo Dataset

The VinDr-Mammo dataset stands as a significant addition to public mammography datasets, offering a detailed resource for breast cancer research. Originating from Vietnam, it encompasses 5,000 mammography exams with thorough annotations, rendering it valuable for AI-based healthcare solutions.

Encompassing various patient demographics and clinical scenarios, the dataset mirrors real-world scenarios encountered in medical practice. It encompasses both screening and diagnostic exams, ensuring its relevance for research.

A notable strength lies in the detailed annotations provided by experienced radiologists, adhering to the BI-RADS standard and encompassing information on breast-level assessments and lesion-level characteristics. This depth of annotation enables detailed analysis and the development of accurate AI models.

The dataset's annotations undergo stringent quality control measures to ensure accuracy and reliability. Additionally, privacy measures are in place to safeguard patient information, aligning with ethical standards.

While VinDr-Mammo offers extensive annotations, it lacks pathology-confirmed ground truth data and essential clinical information like molecular and histology data. Thus, cautious usage is warranted, particularly for CAD systems aimed at

diagnosis. Additionally, limited samples for certain abnormalities impact their reliability for study.

Despite its limitations, VinDr-Mammo represents a valuable resource for developing and evaluating CAD systems for breast cancer screening. Its detailed annotations, including BI-RADS classifications, enable effective training and testing of algorithms. However, users should judiciously use it in conjunction with other clinical data sources.

B. BI-RADS Classification

The Breast Imaging Reporting and Data System (BI-RADS) classification system plays a pivotal role in assessing mammography results, offering detailed insights into breast health status and guiding subsequent medical interventions. Radiologists employ specific criteria to evaluate imaging tests, focusing on various aspects such as masses, breast density, calcifications, asymmetry, and tissue lesions.

Mammograms receive scores based on these criteria, with each BI-RADS category indicating specific findings and suggesting appropriate follow-up care. Categories range from 0 (unclear findings) to 6 (confirmed cancer), encompassing various probabilities and implications for malignancy.

Additionally, mammogram reports often include information about breast density, categorized into four levels (A to D). Higher breast density may pose challenges in detecting changes through mammograms, indicating increased risk for breast cancer.

The BI-RADS classification system serves as a cornerstone in mammography interpretation, providing standardized assessments that guide clinical decision-making and patient management.

C. Swin Transformer Architecture

The Swin Transformer represents a recent breakthrough in the field of computer vision, offering superior performance and scalability compared to previous architectures. Rooted in the Transformer architecture, it introduces innovative concepts tailored for visual data processing.

Hierarchical processing with shifted windows is a key innovation of the Swin Transformer. Unlike traditional CNNs, it dynamically partitions the input image into non-overlapping patches of varying sizes, enabling the model to capture both local and global contextual information effectively. The use of shifted windows enhances the model's ability to maintain spatial relationships across patches, facilitating the capture of long-range dependencies within the image.

Multi-scale feature fusion is another crucial aspect of the Swin Transformer, allowing the model to effectively capture features at various scales. This is achieved through a series of transformer blocks equipped with self-attention mechanisms, facilitating information exchange between patches. Additionally, cross-layer and cross-stage connections enhance feature propagation and gradient flow during training.

Designed to be highly efficient and scalable, the Swin Transformer exhibits linear computational complexity, making it well-suited for both small-scale and large-scale image recognition tasks. Its hierarchical design enables it to handle high-resolution images without significant increase in memory or computational requirements.

Empirical evaluations have demonstrated that the Swin Transformer consistently achieves state-of-the-art performance on various benchmark datasets across a wide range of computer vision tasks. Its hierarchical processing approach, multi-scale feature fusion mechanism, efficient design, and state-of-the-art performance make it a promising choice for advancing the frontiers of visual recognition and understanding.

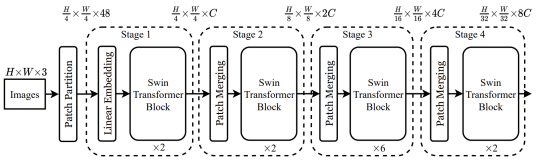

Fig. 1. Swin Transformer Architecture

1) Swin Transformer Architecture: The Swin Transformer architecture, as illustrated in Figure 2, introduces innovative strategies for processing visual data, making it particularly well-suited for tasks such as image classification, object detection, and semantic segmentation.

a) Overall Architecture: The architecture begins by splitting an input RGB image into non-overlapping patches using a patch splitting module, similar to the Vision Transformer (ViT) approach. Each patch is treated as a "token," with its feature represented as a concatenation of the raw pixel RGB values. A linear embedding layer projects these raw-valued features into an arbitrary dimension C.

The Swin Transformer architecture comprises several Transformer blocks, referred to as "Stages," each containing modified self-attention computation (Swin Transformer blocks). These blocks maintain the number of tokens and are responsible for feature transformation at different hierarchical levels.

To produce a hierarchical representation, the number of tokens is gradually reduced through patch merging layers as the network deepens. These merging layers concatenate features from neighboring patches and apply linear transformations to reduce the number of tokens while preserving essential information. The hierarchical representation generated by the Swin Transformer aligns with the feature map resolutions typical of convolutional networks like VGG and ResNet, facilitating seamless integration into existing vision tasks.

b) Swin Transformer Block: The core component of the Swin Transformer architecture is the Swin Transformer block, which replaces the standard multi-head self-attention (MSA) module in traditional Transformer blocks. The Swin Transformer block consists of a shifted window-based MSA module, followed by a two-layer MLP (Multi-Layer Perceptron) with GELU nonlinearity in between. Layer normalization (LN) layers are applied before each MSA module and each MLP, and residual connections are employed to preserve information flow within the block.

c) Shifted Window-based Self-Attention: The standard global self-attention mechanism used in traditional Transformers poses computational challenges for large-scale image processing tasks due to its quadratic complexity with respect to the number of tokens. To address this issue, the Swin Transformer adopts a shifted window-based self-attention approach, which enables efficient computation of self-attention within local windows.

This approach partitions the image into non-overlapping windows and computes self-attention within each window, significantly reducing computational complexity compared to global self-attention. Additionally, to introduce cross-window connections and enhance modeling capabilities, the architecture alternates between two partitioning configurations in consecutive Swin Transformer blocks.

Efficient batch computation techniques, such as cyclic shifting and relative position bias, further enhance the performance and scalability of the shifted window-based self-attention mechanism, making it suitable for a wide range of vision tasks.

d) Architecture Variants: The Swin Transformer architecture offers several variants, each tailored to specific model sizes and computational complexities. These variants include Swin-T, Swin-S, Swin-B, and Swin-L, with varying channel numbers and layer configurations to accommodate different application requirements. These variants enable researchers to choose the most suitable model based on the task's computational constraints and performance objectives.

D. MV-Swin-T

Despite the evident promise of transformers in modeling long-range dependencies, their application in multi-view mammogram analysis remains relatively uncharted territory. Some studies have embraced hybrid models combining transformers and CNNs, introducing global cross-view transformer blocks to amalgamate intermediate feature maps from CC and MLO views. Another noteworthy work is, which employed a transformer-based model for breast cancer segment detection. However, they processed multi-views at a later stage of the network, missing opportunities to capture local correlations between views and lacked results on publicly available datasets, thereby constraining comparability with existing literature.

To fully exploit multi-view insights, a novel transformer-based multi-view network, MV-Swin-T, built upon the Swin Transformer architecture for mammographic image classification is presented. The contributions include designing a novel multi-view network entirely based on the transformer architecture, capitalizing on the benefits of transformer operations for enhanced performance. A novel "Multi-headed Dynamic Attention Block (MDA)" with fixed and shifted window features to enable self and cross-view information fusion from both CC and MLO views of the same breast is introduced. Additionally, the challenge of effectively combining data from multiple views or modalities, especially when images may not align correctly, is addressed. Results using the publicly available CBIS-DDSM And VinDr-Mammo dataset are presented. Moreover, to comprehend the impact of transformers and different associated modules, various architectural changes throughout the training process are introduced, analyzing their overall effects on the entire network.

1) Methodology: The proposed approach focuses on developing a specialized network based on the Swin Transformer architecture, specifically designed for classifying unregistered multi-view mammogram pairs. A novel Omni-Attention transformer block with advanced multi-head dynamic-attention mechanisms and both regular and shifted window configurations is introduced. It aims to effectively integrate multi-view information while addressing challenges related to data alignment and feature fusion.

The Omni-Attention transformer block facilitates self and cross-view information fusion from both CC and MLO views of the same breast. It consists of multiple sub-blocks, including multi-head dynamic attention modules and MLP layers with ReLU non-linearity. Additionally, a shifted window multi-head dynamic attention module is incorporated to enable cross-window connections while preserving computational efficiency.

The proposed MV-Swin-T architecture integrates Omni-Attention blocks into the initial stages of the network, followed by Swin Transformer blocks in subsequent stages. Outputs from different views are concatenated and processed through fully connected layers to maintain consistent dimensions. Finally, the processed output undergoes further transformation to produce the final classification output.

2) Experimental Results: Comprehensive experiments using the CBIS-DDSM and VinDr Mammo datasets were conducted to evaluate the performance of the proposed MV-Swin-T model. Results show significant improvements in classification accuracy compared to traditional single-view approaches. Furthermore, the impact of various architectural changes and transformer modules on the overall performance of the network is analyzed.

III. Methodology

The selection of the Swin Transformer model was driven by several compelling factors. Within the healthcare domain, particularly in fields like mammography, where early detection plays a crucial role, there was a need for an AI solution capable of accurately analyzing complex medical images. The Swin Transformer's innovative design, incorporating self-attention mechanisms and feature fusion, positioned it as an ideal candidate for discerning subtle anomalies indicative of underlying health issues.

Furthermore, the efficiency of the Swin Transformer was particularly appealing. Its ability to handle large datasets like the extensive VinDr mammo dataset without requiring

excessive computational resources made it a promising choice for the project's objectives. However, initial attempts to train the Swin Transformer on the VinDr mammo dataset yielded results below expectations. Despite this setback, the researchers persisted in refining their approach and exploring various adaptations of the model to optimize its performance with the specific dataset.

IV. Experiments and Results

In the initial phases of training with the VinDr Mammo dataset, a range of experiments was conducted using various batch sizes with the Swin Transformer model. The implementation started with a batch size of 32, which yielded an accuracy of approximately 56%. Although this was a decent start, it fell short of expectations. Seeking improvements, the batch size was reduced to 16, resulting in a modest increase in accuracy to around 60%. Encouraged by this progress, the team further reduced the batch size to 8, which surprisingly led to a peak accuracy of 65%.

While these improvements were encouraging, the overall performance was deemed unsatisfactory. The results were compared with the general benchmarks for similar tasks and found to be below the expected standard. Typically, state-of-the-art models in medical image analysis achieve accuracies well above 70%. This realization prompted the exploration of alternative approaches to enhance the model's performance.

In the pursuit of improvement, a variant of the Swin Transformer model was introduced into the training pipeline. This decision proved fruitful, significantly boosting the model's accuracy to an impressive 67%. This was where the model was maxed out, and other methods to the pre-processing were to be explored.

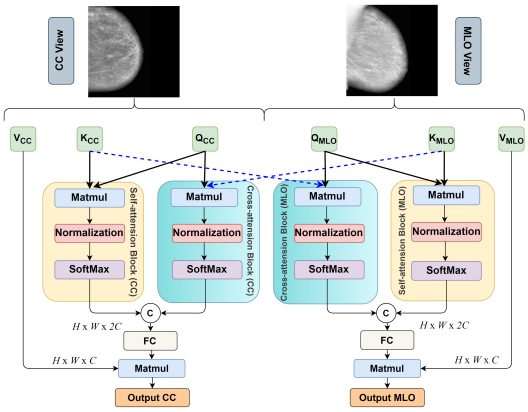

Fig. 2. Swin-MV-T Architecture

V. Future Scope

While the Swin Transformer-based multi-view model (MV-Swin-T) shows promise in various applications and has set benchmarks in several domains, its performance in the context of mammographic image classification seems to be suboptimal. To address this limitation and enhance the model's efficacy, several avenues for future research and development can be explored.

Firstly, integrating preprocessing methods such as windowing into the model pipeline could potentially improve its performance. Windowing involves adjusting the contrast and brightness of medical images to highlight specific anatomical structures or pathologies, thereby enhancing the visibility of relevant features. By incorporating windowing techniques tailored to mammographic images, the model can better capture subtle details and patterns indicative of breast abnormalities, leading to improved classification accuracy.

Additionally, exploring other preprocessing techniques, such as image normalization and enhancement, can further enhance the model's ability to extract relevant features from mammograms. These techniques aim to standardize image characteristics and accentuate diagnostically significant regions, thereby facilitating more accurate classification.

Furthermore, leveraging ensemble learning approaches by integrating multiple transformer and convolutional neural network (CNN) architectures into the model ensemble could lead to significant performance gains. Ensemble methods combine the predictions of multiple models to produce a more robust and accurate classification outcome. By leveraging the complementary strengths of different architectures, the ensemble model can achieve superior performance compared to individual models.

Moreover, given that the primary focus is on achieving benchmark performance regardless of computational efficiency, more extensive model training and parameter tuning can be explored. This includes increasing the model capacity, optimizing hyperparameters, and fine-tuning the training process to maximize classification accuracy.

Overall, by integrating preprocessing methods, exploring ensemble learning approaches, and conducting more extensive model training, the Swin Transformer-based multi-view model can potentially achieve state-of-the-art performance in mammographic image classification.

References

[1] Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, Van Vu, "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," 2023. arXiv:2203.11205 [eess.IV].

[2] Sushmita Sarker, Prithul Sarker, George Bebis, Alireza Tavakkoli, "MV-Swin-T: Mammogram Classification with Multi-View Swin Transformer," February 2024. doi:10.48550/arXiv.2402.16298.