Swin Transformer V2 for Invasive Ductal Carcinoma Classification in Histopathological Imaging

[1]Puguh Aiman Ariyanto, [2*]Untari Novia Wisesty
[1,2]Department of Informatics, Faculty of Informatics, Telkom University, Indonesia
Email: [1]puguhariya@student.telkomuniversity.ac.id, [2]untarinw@telkomuniversity.ac.id

Article Info

Article history:
Received Sep 25th, 2025
Revised Oct 02nd, 2025
Accepted Nov 25th, 2025

Keywords:
Breast Cancer;
Deep Learning;
Histopathological Images;
Invasive Ductal Carcinoma;
Swin Transformer V2

ABSTRACT

Breast cancer is the second leading cause of mortality in women globally, with Invasive Ductal Carcinoma being the most dominant subtype that requires accurate diagnosis to increase patient life expectancy. Conventional diagnosis based on manual histopathological examinations takes a long time, is prone to misinterpretation, and experiences significant inter-observer variability. This study implemented the Swin Transformer V2 architecture for automatic classification of Invasive Ductal Carcinoma on 277,524 histopathological images measuring 50×50 pixels which were resized to 256×256 pixels with geometry augmentation. The model was trained using AdamW optimization with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, batch size 16, and mixed precision FP16 for five epochs at 70:20:10 data sharing. The data augmentation includes a 50% probability of random horizontal flip and a maximum of 10 degrees of random rotation to improve the model's generalization capabilities. Evaluation of 27,754 independent test samples resulted in an accuracy of 92.82%, accuracy of 88.48%, recall of 86.05%, F1-score of 87.25%, and AUC of 0.91. A hierarchical window attention shifted mechanism with residual post-normalization has been shown to be effective in extracting local and global features from complex microscopic images. The results show that Swin Transformer V2 has significant potential as a diagnostic aid system to improve the efficiency and accuracy of early detection of breast cancer in the clinical practice of pathology.

Corresponding Author:
Untari Novia Wisesty
Department of Informatics, Faculty of Informatics, Telkom University
Telekomunikasi Street, Sukapura, Dayeuhkolot District,
Bandung Regency, West Java 40257, Indonesia.
Email: untarinw@telkomuniversity.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.38606

1. INTRODUCTION

Breast cancer is one of the most significant global health problems, especially in female populations around the world. As the most common form of cancer in women, breast cancer ranks second as the leading cause of cancer death after lung cancer, with Invasive Ductal Carcinoma (IDC) is the most dominant subtype and causing worldwide morbidity and mortality in the female population [1]. Current estimates show that by 2024 there will be 310,720 new cases of breast cancer diagnosed in the United States, with a mortality rate of about 42,250 women dying from the disease [2].

In conventional clinical practice, physicians and pathologists often use histopathological images to analyze breast tumor tissue that is then classified as benign (Benign) or violent (malignant) [3]. This manual examination process has a number of fundamental limitations that need to be overcome in the context of improving the quality of modern diagnostic services. First, the procedure of visual interpretation of histopathological images requires a substantial duration of time due to the complexity of cellular and tissue structures that must be evaluated in detail by the pathologist. Second, manual examinations by pathologists can be tiring and prone to errors and take a considerable amount of time in the process. Third, the variability of interpretation between observers (Inter-observer variability) is a challenge in maintaining the standardization of pathology diagnosis in various health institutions. Early and accurate diagnosis is essential to improve the patient's life expectancy through timely therapeutic interventions [4].

Technological developments Deep Learning In the last decade, it has revolutionized many fields in healthcare for various tasks, such as accurate diagnosis and prognosis of diseases, as well as robot-assisted surgeries. Rapid advances in artificial neural network architecture have enabled the development of automated systems for the classification of medical images that can save time and reduce misdiagnosis [5]. Technology Deep Learning has demonstrated an outstanding ability to extract complex features from histopathological images that are difficult to identify manually, thus opening up great opportunities to improve accuracy and efficiency in breast cancer diagnosis through more objective and consistent analysis of medical images [6].

One of the significant breakthroughs in architecture Deep Learning is the emergence of models Transformers which was originally developed for natural language processing but was later adapted for computer vision. Swin Transformer is a variant Vision Transform optimized to work with high-resolution images such as histopathological images, using Shifted Windows which improves the efficiency of calculations Self-attention while maintaining the model's ability to capture local and global features in the image [7] Previous research has shown that the model is based on Swin Transformer can achieve excellent results in a wide range of medical image classification tasks, with an accuracy of up to 99.6% for the classification of two classes (Benign Versus malignant) and 96.0% for the classification of eight breast cancer subtypes using the BreaKHis dataset [5].

Even though Swin Transformer The first generation has demonstrated impressive performance, but it still faces challenges in terms of computational efficiency and adaptation to variations in feature scale in complex histopathological images. Development Swin Transformer V2 It is present as a solution to overcome these limitations with various significant architectural improvements. Swin Transformer V2 Introducing the technique Residual Post-Normalization To improve training stability, the Scaled cosine attention to improve model transferability between different image resolutions, as well as strategies log-spaced continuous position bias which allows the model to better adapt to variations in input sizes. These innovations make it possible for Swin Transformer V2 to achieve a larger model capacity of up to 3 billion parameters and an image resolution of up to 1536×1536 pixels, which is particularly relevant for detailed analysis of histopathological images that require feature detection capabilities at various spatial scales [4].

This study proposes the application of Swin Transformer V2 to a different dataset from the previous study, namely the Kaggle Predict IDC in Breast Cancer Histology Images competition dataset which specifically focuses on the classification of IDC (Predict IDC in Breast Cancer Histology Images, 2025). This dataset has the unique characteristic of a 50×50-pixel histopathological image patch with IDC and non-IDC binary labels, which demands the model's ability to capture microscopic details at very small scales. The approach proposed in this study involves a systematic hyperparameter fine-tuning process to optimize the performance of the Swin Transformer V2 on the dataset, including learning rate adjustments, batch size, data augmentation strategies, and architectural configurations tailored to the characteristics of breast cancer histopathological data.

The innovative value of this study lies in the in-depth exploration of the ability of Swin Transformer V2 to classify IDC in small histopathological image patches, which differs from most previous studies that used higher resolution images or datasets with different characteristics. This research also contributes in the form of a comprehensive analysis of effective Fine-tuning strategies for Transformers models in the histopathological imaging domain, which can be a reference for similar studies in the future. In addition, this study is expected to produce a faster, more accurate, and reliable artificial intelligence-based diagnostic aid system in diagnosing IDC, so as to increase the efficiency of the clinical diagnosis process and ultimately contribute to improving the quality of health services and life expectancy of breast cancer patients through more optimal early detection [8].

Previous research has shown different approaches to the classification of breast cancer using deep learning. A comprehensive survey of Transformer's applications in medical imaging found that attention-mechanism-based architectures are capable of capturing long-term spatial dependencies that conventional CNN cannot achieve, but most studies are still focused on standard Vision Transformers that have quadratic computational complexity to image resolution [9]. TransUNet combines CNN with Transformer for medical image segmentation, but this hybrid architecture requires two separate backbones that significantly improve model parameters [10]. In the histopathology domain of specific breast cancer, a pure CNN approach was used for IDC classification with 89% accuracy, but CNN's limitations in capturing a global context caused the model to fail to detect invasive patterns spread across the network [11]. Addressing this through the CNN ensemble method with undersampling techniques to address class imbalance, this approach risks removing important information from the majority class and resulting in an F1-score of only 85% [12].

The novelty of this research lies in the application of the Swin Transformer V2 architecture which is specifically designed to overcome the limitations of the previous models through three main innovations: (1) Residual Post-Normalization which improves training stability in deep tissues without the need for additional normalization techniques, (2) Scaled Cosine Attention which makes the model robust to the contrast variations of histopathological staining without relying on extensive preprocessing, and (3) Log-spaced Continuous Position Bias that allows effective transfer of learning from ImageNet to different medical image resolutions. In contrast to the study that used Swin Transformer V1 on the high-resolution BreaKHis dataset, this study explored the capabilities of Swin V2 on the IDC Kaggle dataset which has the unique characteristic of a very small image patch (50×50 pixels) that is then resized to 256×256 pixels, a condition that has never been comprehensively evaluated in the literature before and demands feature extraction capabilities at extreme microscopic scales [13].

2. RESEARCH METHOD

This study implements the Swin Transformer V2 architecture for the classification of IDC in histopathological imaging of breast cancer with a supervised machine learning approach. The dataset used came from a Kaggle competition titled Predict IDC in Breast Cancer Histology Images which contained 277,524 50×50 pixel image pieces with three RGB color channels extracted from 162 Hematoxylin and Eosin (H&E) colored histopathology slides at 40x microscopic magnification (Predict IDC in Breast Cancer Histology Images, 2025). The dataset consisted of two diagnostic categories: 78,786 IDC-positive images and 198,738 non-IDC-negative labeled images, with each image file following a naming convention that included the patient's identity, the spatial coordinates of the piece, and a binary class label (0 for non-IDC and 1 for IDC) (Predict IDC in Breast Cancer Histology Images, 2025).

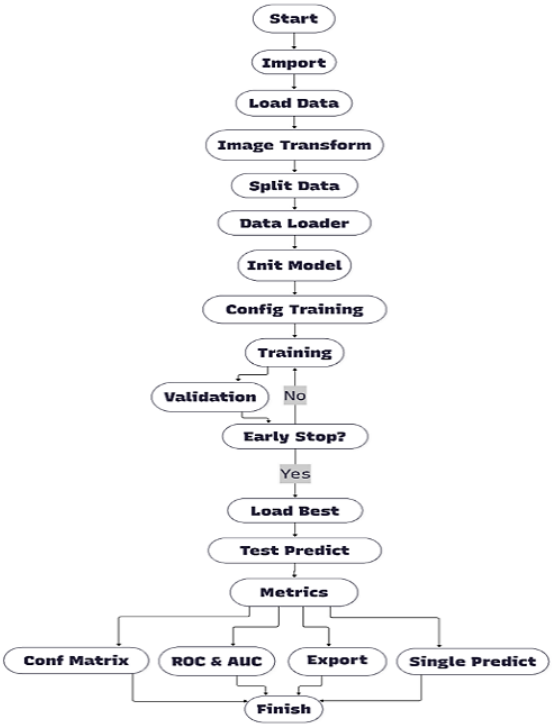

Figure 1. Classification System Workflow

The system workflow is designed following standard stages Machine Learning Pipeline starting from data acquisition to model performance evaluation as shown in Figure 1. The pre-processing stage involves geometry transformation and pixel normalization using the library torchvision.transforms, where the image is re-resolved to 256×256 pixels to match the input dimensions to the architecture pre-trained Swin Transformer V2. In the training data, augmentation is applied in the form of Random Horizontal Flip with a probability of 50%, random rotation up to 10 degrees, conversion to the PyTorch tensor, as well as normalization using the value Mean [0.485, 0.456, 0.406] and Standard Deviation [0.229, 0.224, 0.225] as per ImageNet standards. Meanwhile, validation and test data only underwent a process Resize and normalization without random augmentation to maintain evaluation consistency. The dataset sharing is carried out using the Random split with a proportion of 70% for training (194,266 samples), 20% for validation (55,504 samples), and 10% for testing (27,754 samples). This sharing strategy aims to ensure the model obtains adequate training data while providing a separate validation set for monitoring the performance of each Epoch and prevent Overfitting through the mechanism Early Stopping [14].

The implemented model is a variant swinv2_base_window8_256 who have gone through Pre-training on ImageNet-1K, with major architectural modifications in the form of Residual Post-Normalization for training stability on the inner network, Scaled cosine attention that replaces the mechanism Dot Product Attention standards to be more robust against histopathological staining contrast variations, and log-spaced continuous position bias which allows effective transferability between different resolutions.

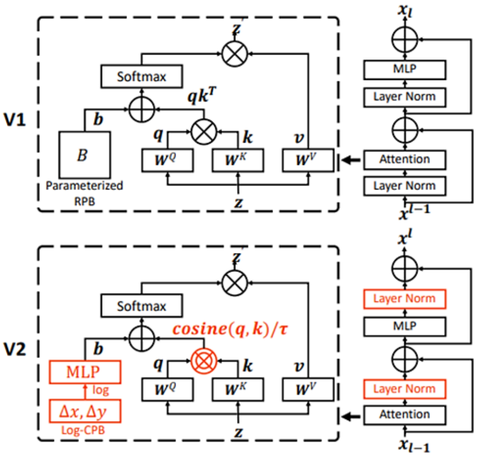

Figure 2. Architectural Comparison between Swin Transformer V1 and Swin Transformer V2

The structural differences between Swin Transformer V1 and V2 are illustrated in Figure 2, where V2 introduces three key modifications to enhance model capacity and resolution scalability.
Mathematical Formulation of Swin Transformer V2 Architecture
1. Residual Post-Normalization

Swin V2 changes the position of the Normalization Layer from before (Pre-Norm) to after (Post-Norm) residual connection:

$$\text{Swin V1 (Pre-Standard): } x_{l+1} = x_l + \text{Block}(\text{LN}(x_l)) \tag{1}$$

$$\text{Swin V2 (Post-Standard): } x_{l+1} = \text{LN}(x_l + \text{Block}(x_l)) \tag{2}$$

This Post-Norm configuration stabilizes the activation distribution in deep tissues, preventing the gradient exploding that often occurs when training models with histopathological images that have high contrast variations between slides [13].

2. Scaled Cosine Attention

The dot-product() based attention mechanism is replaced by normalized cosine similarity: qk^T

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{\cos\,(Q,K)}{\tau} + B\right) \cdot V \tag{3}$$

with and as a learnable temperature parameter. This approach makes the calculation of attention independent of the magnitude of the pixel value, so that the model is robust against variations in Hematoxylin-Eosin staining intensity in different histopathological preparations [7].

$$\cos\,(Q, K) = \frac{Q\cdot K^T}{\|Q\|\|K\|}\,\tau \tag{4}$$

3. Log-spaced Continuous Position Bias

For flexible window resolution adaptation, relative position bias is defined in logarithmic space:

$$\widehat{\Delta}x = \text{sign}(\Delta x) \cdot \log\,(1 + |\,\Delta x\,|) \tag{5}$$

$$\widehat{\Delta}y = \text{sign}(\Delta y) \cdot \log\left(1 + |\Delta y|\right) \tag{6}$$

The final bias is generated via the Multi-Layer Perceptron:

$$B(\Delta x, \Delta y) = G(\widehat{\Delta}x, \widehat{\Delta}y) \tag{7}$$

This logarithmic transformation reduces the extrapolation ratio by up to 60% when models trained at 8×8 window sizes are transferred to different resolutions, allowing for efficient fine-tuning of pretrained ImageNet weights without significant performance degradation in the histopathological domain [7].

Hyperparameter configurations include AdamW optimization with Learning Rate 1×10^{-4} then Weight Decay 1×10^{-4}, loss function cross entropy loss, as well as scheduling Learning Rate adaptive using ReduceLROnPlateau with a factor of 0.5 and Patience 2. Training is carried out for a maximum of 5 epochs, with early stopping patience set to 5 epochs, where the model with the lowest validation loss is kept as the best model. Performance evaluation is performed on a set of tests that have never been seen during training using accuracy metrics, Accuracy, Recall, F1 Score, a confusion matrix, as well as an ROC curve with an AUC value to measure the discriminatory ability of the model at various probability thresholds [14].

.

3. RESULTS AND ANALYSIS

3.1. Results and Analysis

3.1.1. Experimental Environment and System Configuration

The implementation of the Swin Transformer V2 architecture for classifying IDC in histopathological images of breast cancer was carried out using a GPU-based computing infrastructure with the support of CUDA acceleration to optimize the training speed of complex models. Experiments were conducted in the Jupyter Notebook environment using Intel Core i7/Xeon processor specifications, NVIDIA GPUs with FP16 mixed precision capability, 16-32 GB of RAM, and SSD storage integrated with cloud storage. Software configurations include Python 3.10 as the programming language, PyTorch 2.x as the primary deep learning framework, TIMM libraries for access to pre-trained models, as well as supporting libraries such as Torchvision for image processing, Scikit-learn for model evaluation, and Matplotlib and Seaborn for visualizing analysis results. The dataset used consists of approximately 277,000 pieces of PNG-formatted histopathology images with RGB color channels grouped by patient identity, with the image resolution standardized to 256×256 pixels to meet the input dimension requirements of the Swin Transformer V2 architecture.

3.1.2. Pre-Processing and Dataset Sharing

The pre-processing stage involves geometry transformation and pixel intensity normalization, where the training data is applied an augmentation technique in the form of a random horizontal flip with a probability of 50%, random rotation of a maximum of 10 degrees, and normalization using ImageNet standard parameters with mean values [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] to improve the model's generalization ability against the orientation variation and staining characteristics of histopathological tissues. The distribution of the dataset was carried out with a ratio of 70:20:10 for the set of training, validation, and tests, resulting in a distribution of 194,266 samples for training, 55,504 samples for validation, and 27,754 samples for final testing (see Table 1). This sharing strategy is designed to ensure the model obtains adequate training data while providing a separate validation set for monitoring the performance of each epoch and preventing overfitting through an early stoppings mechanism.

Table 1. Sample Description

Dataset Types	Percentage	Number of Samples (Estimated)
Training Set	70%	194,266 samples
Validation Set	20%	55,504 samples
Testing Set	10%	27,754 samples

3.1.3. Model Configuration and Training Process

The model used is a swinv2_base_window8_256 variant with a pre-trained initialization weight from ImageNet, which then adjusts the output layer to two neurons for the binary classification of benign and malignant classes using the cross-entropy loss function. The training hyperparameter configuration includes the use of AdamW optimization with an initial learning rate of 1×10^{-4} and a decay weight of 1×10^{-4}, a batch size of 16 samples, and adaptive learning rate scheduling via ReduceLROnPlateau which automatically lowers the learning rate value by 50% when validation performance stagnates for two epochs consecutively.

The training was conducted over five epochs with the application of FP16 mixed precision to improve computing efficiency and reduce GPU memory consumption, with training speeds recorded at about one to two minutes for every 3,000 batches. Hyperparameter can be seen in Table 2.

Table 2. Hyperparameter

Hyperparameter	Value
Learning Rate	1e-4 (with ReduceLROnPlateau)
Optimizer	AdamW (Weight Decay 1e-4)
Loss Function	CrossEntropyLoss
Batch Size	16
Number of Epochs	5
Accuracy	Mixed Precision (FP16)

The training process exhibited progressive and stable convergence, with the model achieving a validation accuracy of 91.70% in the first epoch, along with a training loss of 0.2543 and a validation loss of 0.2287. The improvement in performance continued until the fifth epoch, where the training accuracy reached 92.88% with a loss of 0.2010, while the validation accuracy reached 92.83% with a loss of 0.1961. The very small difference in loss between the training and validation sets indicates that the model does not experience significant overfitting, but rather achieves an optimal balance between learning capabilities and generalizations to new data.

3.1.4. Evaluation Results on Test Data

Final performance evaluation is performed using a set of tests that the model never saw during the training process. The resulting confusion matrix showed that out of a total of 27,754 test samples, the model successfully classified 18,939 benign samples correctly (true negative), while 888 benign samples were incorrectly predicted as malignant (false positive). In the malignant class, the model managed to identify 6,821 samples correctly (true positive), but there were 1,106 samples that failed to be detected and predicted as benign (false negative). This prediction distribution indicates that the model has good sensitivity in detecting non-cancerous samples, but still has the challenge of minimizing prediction errors in cancer classes which is a critical aspect in medical diagnostic applications. The convergence pattern during the training process is visualized in Figure 3, which illustrates the progression of training and validation accuracy alongside their corresponding loss values across five epochs.

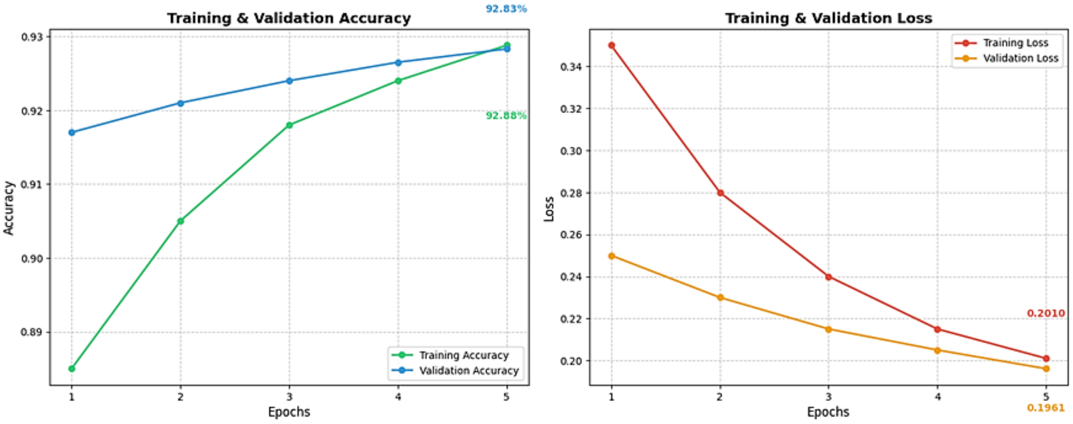

Figure 3. Training Graph

Figure 3 demonstrates a progressive and stable convergence pattern where both training and validation curves follow parallel trajectories without significant divergence, indicating that the model successfully learned generalizable features from the histopathological images rather than memorizing training samples. The validation accuracy increased from 91.70% in the first epoch to 92.83% in the fifth epoch, while the validation loss decreased from 0.2287 to 0.1961, suggesting that the applied augmentation techniques and adaptive learning rate scheduling effectively prevented overfitting and enabled the model to capture robust discriminative patterns in breast cancer tissue microstructures.

3.1.5. Error Pattern Analysis and Clinical Implications

Detailed examination of the 1,106 false negative cases reveals systematic error patterns that warrant further investigation. The misclassified samples predominantly exhibited ambiguous morphological characteristics where the boundaries between benign ductal epithelium and early invasive carcinoma cells were not clearly demarcated, presenting transitional features that challenge even experienced pathologists during manual examination. Additionally, the upscaling process from original 50×50 pixel patches to 256×256 pixels through bilinear interpolation may have introduced smoothing artifacts that obscured fine-grained nuclear texture details critical for detecting subtle invasive patterns, particularly in cases where malignant cells infiltrate the stroma in scattered individual cell formations rather than cohesive tumor nests. The 888 false positive predictions, while less clinically critical than false negatives, primarily occurred in tissue regions containing inflammatory cell infiltrates, fibrotic stroma with reactive epithelial changes, or areas with severe compression artifacts from tissue processing, which generated visual patterns that mimicked the hypercellularity and nuclear pleomorphism characteristic of invasive carcinoma. These error patterns suggest that future model improvements should focus on incorporating multi-scale feature extraction strategies to preserve fine morphological details and implementing attention-weighted loss functions that penalize false negatives more heavily than false positives to align the model's optimization objective with clinical safety priorities in cancer screening applications. Figure 4 show confusion matrix prediction results in data testing.

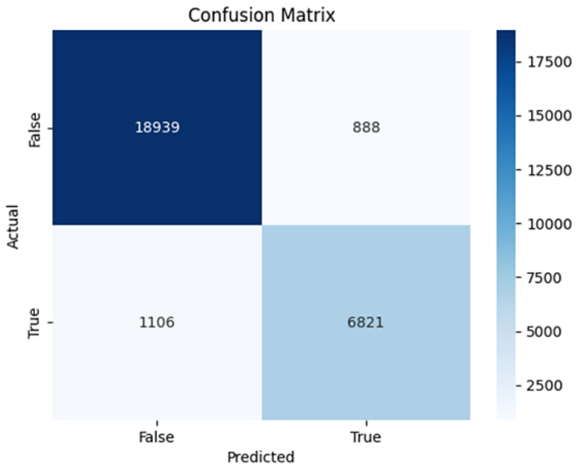

Figure 4. Confusion Matrix Prediction Results in Data Testing

Analysis of the Receiver Operating Characteristic (ROC) curve in Figure 5 showed that the model has excellent discriminating ability between benign and malignant classes, with the curve moving well above the random diagonal line. The Area Under Curve (AUC) value obtained reached 0.91, indicating that the model was able to distinguish the two classes with a high level of confidence at various classification probability thresholds. The higher the AUC value, the better the model's ability to reduce misclassification, especially in reducing false negatives in cancer detection cases that can have a serious impact on the patient's clinical management.

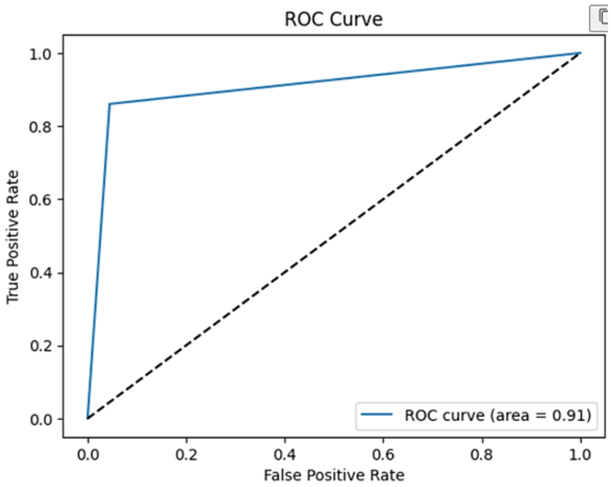

Figure 5. ROC Curve

3.1.6. Classification Performance Metrics

A summary of the standard evaluation metrics in the test set shows the model's highly competitive performance (see Table 3). The overall accuracy reached 92.82%, indicating that the majority of the model's predictions corresponded to the actual labels for both diagnostic classes. A precision value of 88.48% indicates a good level of accuracy in predicting the positive class of cancer, so that the risk of healthy patients being classified as cancer patients is within acceptable limits for initial screening applications. The recall value or sensitivity reached 86.05%, indicating that the model is able to detect the majority of cancer cases that are actually positive, although there is still room for improvement to minimize false negative cases that can result in delays in diagnosis and treatment. The F1-score of 87.25% confirms the harmonious balance between precision and recall, showing that the model's performance does not lean towards just one aspect but achieves the optimal trade-off between precision and sensitivity.

Table 3. Model Evaluation Results on Data Testing

Metric	Value	Interpretation
Accuracy	0.9282 (92.82%)	The overall prediction accuracy rate is very high.
Accuracy	0.8848 (88.48%)	The accuracy of positive (malignant) predictions is quite reliable.
Recall	0.8605 (86.05%)	The sensitivity of the model in detecting breast cancer is good.
F1 Score	0.8725 (87.25%)	A harmonious balance between Precision and Recall.

The evaluation results in Table 3 show that the Swin Transformer V2 has excellent generalization capabilities on test data that has never been seen before. The distribution of the image samples that were successfully classified appropriately showed that the model has a robust ability to recognize complex histopathological visual patterns. Correctly predicted benign samples exhibited the characteristics of an organized tissue structure with a normal cellular architecture, while accurately identified malignant samples exhibited abnormal carcinoma cell infiltration characteristics with invasive growth patterns. Some cases of prediction errors occur mainly in samples with ambiguous characteristics where the boundaries between normal and pathological tissues are not very clear, or in image pieces that contain staining artifacts that can affect the extraction of features by the model.

3.2. Discussion

3.2.1. Comparison of Performance with Previous Research

Performance model Swin Transformer V2, which achieved 92.82% accuracy on the classification IDC, shows very competitive results and is comparable to previous studies using architecture Deep Learning for histopathological image analysis of breast cancer. Comparison with previous research reveals that traditional CNN-based convolutional architectures have been widely applied for classifying histopathological images with varying degrees of accuracy, depending on the complexity of the architecture and the training strategies employed (Islam et al., 2024). Studies using the CNN ensemble approach for the classification of breast cancer histological images reported fairly good performance with an accuracy ranging from 85-90% on similar datasets, but with higher computational requirements because they involved a combination of multiple models [15]. A hybrid approach that combines the CNN architecture with the mechanism of Transformers has also been explored and shown an increase in accuracy of up to 94% in some cases, but with higher implementation complexity and longer training times [16].

Superiority Swin Transformer V2 lies in its ability to capture long-distance dependencies between image regions through a mechanism of Self-attention global, while maintaining computing efficiency through a Shifted Window that limits the calculation to local overlapping windows. Application Deep Learning Histopathological imaging of breast cancer has shown great potential in improving diagnostic accuracy, accelerating the pathology analysis process, and providing a more accurate prognosis to support better clinical decisions [17]. Recent developments in this field show that architecture-based Transformers It has a superior ability to handle complex variations in medical imagery compared to traditional convolutional architectures, which have limitations in capturing the global context in high-resolution images [18].

3.2.2. Prediction Error Analysis and Clinical Implications

Analysis of the fusion matrix revealed that the model had a higher tendency to produce false negatives (1,106 cases) than false positives (888 cases). This phenomenon has significant clinical implications because a false Negative means that a truly positive cancer case fails to be detected by the system, which can result in delayed diagnosis and delays in necessary therapeutic interventions. From the perspective of public health and clinical management, this type of error is more dangerous than false positives, which, although they can cause patient anxiety and unnecessary follow-up examinations, do not result in the risk of untreated disease progression. The epidemiology of breast cancer shows that early and accurate detection is crucial to improve cure rates and reduce mortality, so minimizing false negatives should be a top priority in the development of artificial intelligence-based diagnostic systems for medical applications [12].

Several factors can explain the occurrence of prediction errors in the model. First, the intrinsic characteristics of the dataset that uses very small pieces of imagery (50×50 pixels) which are then analyzed.Resize Being 256×256 pixels can result in the loss of important microscopic details or the appearance of interpolation artifacts that affect feature extraction by neural networks. Second, variability in the preparation and staining techniques of inter-laboratory histopathological tissue can result in differences in color intensity and contrast that affect the consistency of visual representations, with studies suggesting that age and patient sex factors may also correlate with the visual characteristics of the molecular subtypes of breast cancer [19]. Third, the biological complexity of the transition ductal carcinoma in situ To IDC involving a spectrum of gradual morphological changes can result in ambiguous samples that are difficult to classify, with complex molecular features and clinical significance in the process of disease progression [20].

3.2.3. Evaluation of Diagnostic Metrics and Clinical Applications

Value Accuracy of 88.48% indicates that when the model predicts a sample as malignant, there is a probability of around 88.48% that the prediction is correct. This metric is important in the clinical context to avoid overdiagnosis that can result in unnecessary invasive diagnostic procedures, patient psychological anxiety, as well as an increased burden of healthcare costs. Meanwhile, the value of Recall 86.05% indicated that of all cancer cases that were actually positive, the model was able to detect about 86.05% of them. Increased model sensitivity can be achieved through more conservative adjustment of the classification probability threshold, although this will lower Accuracy and improve False positive axle Trade-off which should be considered in the context of clinical application and acceptable risk preferences [14].

Use Transfer Learning with weights Pre-trained from ImageNet has proven to be effective in accelerating convergence and achieving high performance with a number of Epoch limited training. However, the fundamental question regarding Domain Gap The relationship between ImageNet natural imagery and medical histopathology imagery remains relevant to discuss. Application Deep Learning Histopathological imaging is not only limited to diagnosis, but also includes prediction of response to therapy and estimation of the patient's long-term prognosis, so the development of robust and reliable models is critical to support comprehensive clinical decision-making [20]. Research shows that although low-level representations of features such as edges, textures, and color gradients can be transferred across domains, high-level semantic features specific to pathological interpretation may require more extensive retraining or the use of datasets Pre-training which is more relevant to the medical domain [10].

3.2.4. Research Limitations and Development Recommendations

The limitations of this study include the use of a single dataset that, although large in size, comes from a homogeneous patient cohort and acquisition protocol, so that the model's generalizability to inter-institutional variation, demographic differences, and variation in network preparation techniques cannot be fully ascertained. Advanced research with external validation using independent datasets from various sources is indispensable to test robustness and model reliability in various clinical conditions [21]. In addition, the interpretability of model predictions through visualization techniques such as Gradient-weighted class activation mapping can provide insight into which areas of imaging have the most influence on classification decisions, thereby increasing clinicians' confidence in artificial intelligence-based systems and facilitating the process of verifying diagnoses by pathologists [6].

This study acknowledges several methodological limitations that provide opportunities for future research directions. First, the reliance on a single-institution dataset from the Kaggle competition, despite its substantial size of 277,524 image patches, limits the generalizability of the model to inter-institutional variations in tissue preparation protocols, scanner characteristics, and demographic diversity of patient populations, necessitating external validation using independent multi-center datasets to establish clinical robustness across diverse healthcare settings [22]. Second, the binary classification framework focusing exclusively on IDC versus non-IDC does not capture the full spectrum of breast cancer molecular subtypes and histological grades that influence treatment planning, suggesting that future iterations should incorporate multi-class classification capabilities to distinguish between luminal A, luminal B, HER2-enriched, and triple-negative subtypes based on integrated histopathological and immunohistochemical features [5]. Third, the current implementation treats each 50×50 pixel patch as an independent diagnostic unit without considering spatial relationships between adjacent tissue regions, whereas recent advances in graph neural networks and whole-slide image analysis have demonstrated that contextual information from surrounding tissue architecture significantly improves diagnostic accuracy by capturing tumor-stroma interactions and growth patterns at multiple spatial scales [23].

Fourth, the black-box nature of deep learning models poses challenges for clinical adoption, as pathologists require interpretable explanations of model predictions to verify diagnostic reasoning and identify potential failure modes, highlighting the need for integrating explainable AI techniques such as attention visualization, feature attribution maps, and counterfactual explanations that align with established histopathological criteria used in human expert diagnosis. Finally, the computational requirements of Swin Transformer V2, which necessitate GPU acceleration and mixed precision training, may limit deployment in resource-constrained clinical laboratories, suggesting that future research should explore model compression techniques, including knowledge distillation, neural architecture search, and quantization-aware training to develop lightweight variants that maintain diagnostic performance while enabling real-time inference on standard pathology Workstations [24].

The practical implications of this study show that Swin Transformer V2 It has the potential to be implemented as an automated triage system that can prioritize high-probability cases for review by pathologists first, thereby improving the efficiency of laboratory workflows and speeding up diagnosis time. The integration of this system in Digital Pathology Workflow can also serve as a Second opinion, which is objective to reduce inter-observer variability and improve the consistency of pathological diagnosis in various health institutions. However, clinical implementation still requires comprehensive prospective validation, regulatory approval from the competent health authority, evaluation of the impact on the Clinical Outcome of patients, as well as ethical and legal considerations before they can be widely adopted in the standard practice of breast cancer pathological diagnosis [25].

4. CONCLUSION

This study successfully implemented the Swin Transformer V2 architecture for IDC classification on breast cancer histopathological images, achieving an accuracy of 92.82%, a precision of 88.48%, a recall of 86.05%, an F1-score of 87.25%, and an AUC of 0.91 on a dataset of 277,524 image patches. The hierarchical shifted window attention mechanism proved effective in extracting complex features from microscopic images, while the hyperparameter fine-tuning strategy, combined with AdamW optimization and an adaptive learning rate scheduler, resulted in stable convergence without significant overfitting. Limitations of this study include: First, reliance on a single-institution dataset from Kaggle limits generalizability to inter-institutional variations in tissue preparation protocols, scanner characteristics, and demographic diversity, necessitating external validation using independent multi-center datasets. Second, the binary classification framework (IDC vs. non-IDC) does not capture the full spectrum of breast cancer molecular subtypes that influence treatment planning. Third, treating each 50×50 pixel patch independently, without considering the spatial relationships between adjacent tissue regions, may overlook contextual information that could improve diagnostic accuracy.

Fourth, the black-box nature of deep learning models poses challenges for clinical adoption, as pathologists require interpretable explanations of predictions. Finally, computational requirements necessitating GPU acceleration may limit deployment in resource-constrained laboratories. Future research should focus on: implementing class imbalance handling techniques (SMOTE, Focal Loss), integrating explainable AI methods (Grad-CAM, attention visualization), exploring multi-class classification for molecular subtypes, developing graph neural networks for whole-slide analysis, and applying model compression techniques for real-time inference on standard pathology workstations.

REFERENCES

[1] A. Ibrahim, H. Torkey, and A. El-Sayed, "Invasive Ductal Carcinoma (IDC) Nuclei Classification using Mask RCNN," *Menoufia J. Electron. Eng. Res.*, vol. 34, no. 2, pp. 20–30, 2025, doi: 10.21608/mjeer.2025.351724.1103.

[2] A. Zeynali, M. A. Tinati, and B. M. Tazehkand, "Hybrid CNN-Transformer Architecture With Xception-Based Feature Enhancement for Accurate Breast Cancer Classification," *IEEE Access*, vol. 12, no. December, pp. 189477–189493, 2024, doi: 10.1109/ACCESS.2024.3516535.

[3] M. Thahiruddin, "CNNs vs. Hybrid Transformers for Brain Tumor Classification on the BRISC Dataset," *J. Apl. Technology. Inf. and Manaj.*, vol. 6, no. 1, pp. 24–33, 2025, doi: 10.31102/jatim.v6i1.3545.

[4] O. Tanimola, O. Shobayo, O. Popoola, and O. Okoyeigbo, "Breast Cancer Classification Using Fine-Tuned SWIN Transformer Model on Mammographic Images," *Analytics*, vol. 3, no. 4, pp. 461–475, 2024, doi: 10.3390/analytics3040026.

[5] S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers," *Mathematics*, vol. 10, no. 21, 2022, doi: 10.3390/math10214109.

[6] B. Jiang, L. Bao, S. He, X. Chen, Z. Jin, and Y. Ye, "Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis," *Breast Cancer Res.*, vol. 26, no. 1, 2024, doi: 10.1186/s13058-024-01895-6.

[7] Z. Liu *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," *Proc. IEEE Comput. Soc. Conf.*

Comput. Vis. Pattern Recognit., vol. 2022-June, pp. 11999–12009, 2022, doi: 10.1109/CVPR52688.2022.01170.

[8] B. Guo, X. Li, M. Yang, J. Jonnagaddala, H. Zhang, and X. S. Xu, "Predicting microsatellite instability and key biomarkers in colorectal cancer from H&E-stained images : achieving state-of-the-art predictive performance with fewer data using Swin Transformer," no. May, pp. 223–235, 2023, doi: 10.1002/cjp2.312.

[9] J. Saeed and M. Hussein, "Empowering Ovarian Cancer Subtype Classification with Parallel Swin Transformers and WSI Imaging," vol. 21, no. 6, pp. 1006–1014, 2024.

[10] V. Sreelekshmi, K. Pavithran, and J. J. Nair, "SwinCNN : An Integrated Swin Transformer and CNN for Improved Breast Cancer Grade Classification," *IEEE Access*, vol. 12, no. May, pp. 68697–68710, 2024, doi: 10.1109/ACCESS.2024.3397667.

[11] S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers," *Mathematics*, vol. 10, no. 21, pp. 1–7, 2022, doi: 10.3390/math10214109.

[12] B. Yuan, Y. Hu, Y. Liang, Y. Zhu, and L. Zhang, "Comparative analysis of convolutional neural networks and transformer architectures for breast cancer histopathological image classification," vol. 999, no. June, pp. 1–17, 2025, doi: 10.3389/fmed.2025.1606336.

[13] Y. Vishwakarma and A. A. Waoo, "Swin Transformer for Breast Cancer Classification using Histopathology Images," vol. 11, no. 3, pp. 120–127, 2023.

[14] A. Stanisławek, "Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review," pp. 1–30, 2021.

[15] G. Aji Mahesa, "Classification of Histological Images of Breast Cancer Using the CNN Ensemble Method," *J. Repos.*, vol. 4, no. 3, pp. 373–384, 2022, doi: 10.22219/repositor.v4i3.1497.

[16] H. C. Tang, Y. Dong, T. Huang, M. F. Han, and J. Fu, "Foreign object detection for transmission lines based on Swin Transformer V2 and YOLOX," *Vis. Comput.*, vol. 40, no. 5, pp. 3003–3021, 2024.

[17] P. Examination, "Utilization of Deep Neural Networks in tasks of classification and semantic segmentation of medical images of colon and breast cancer," no. January, pp. 1–3, 2024.

[18] B. Yang *et al.*, "Transformer-based multiple instance learning network with 2D positional encoding for histopathology image classification Introduction," *Complex Intell. Syst.*, vol. 11, no. 5, pp. 1–17, 2025, doi: 10.1007/s40747-025-01779-y.

[19] J. Wang *et al.*, "Progression from ductal carcinoma in situ to invasive breast cancer: molecular features and clinical significance," *Signal Transduct. Target. Ther.*, vol. 9, no. 1, 2024, doi: 10.1038/s41392-024-01779-3.

[20] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," pp. 1–11, 2021.

[21] M. A. Rahmadi, H. Nasution, L. Mawar, N. Sihombing, R. Nasution, and M. Sari, "The Effect of Anxiety on Breast Cancer Treatment Adherence," *J. Ilm. Health Sciences. and Medicine.*, vol. 3, no. 1, 2025.

[22] D. S. Dizon and A. H. Kamal, "Cancer statistics 2024: All hands on deck," *CA Cancer J. Clin*, vol. 74, no. 1, 2024, doi: 10.3322/caac.21824.

[23] Breastcancer.org, "Your Guide to the Breast Pathology Repor," 2024.

[24] S. S. Kumar, "Advancements in medical image segmentation: A review of transformer models," *Comput. Electr. Eng.*, vol. 123, 2025, doi: 10.1016/j.compeleceng.2025.110099.

[25] S. Liu and B. Min, "DCS-ST for Classification of Breast Cancer Histopathology Images with Limited Annotations," *Appl. Sci.*, vol. 15, no. 15, pp. 1–9, 2025, doi: 10.3390/app15158457.

BIBLIOGRAPHY OF AUTHORS

Dr. Untari Novia Wisesty earned her S.T. and M.T. degrees in Informatics Engineering from Telkom Institute of Technology (now Telkom University), then completed her Doctorate in Electrical Engineering and Informatics at the Bandung Institute of Technology with a dissertation on IM_SelaTCN deep learning models for cancer DNA sequence labeling. He is a lecturer and researcher in the Data Science and Intelligent Systems Expertise Group at the Faculty of Informatics, Telkom University, with expertise in the fields of machine learning, artificial intelligence, bioinformatics, and Brain-Computer Interfaces. His research includes the development of PCA methods for cancer detection from microarray data, as well as various machine learning studies for biomedical data. In addition, Dr. Untari actively manages scientific journals, including as Editor-in-Chief of JASMINE: Journal of Intelligent Systems and Machine Learning, thereby contributing to the improvement of publication quality in the field of artificial intelligence.

Puguh Aiman Ariyanto is an S1 student of the Informatics study program class of 2022 at Telkom University, Bandung. He is currently completing a final project focusing on the application of Deep Learning and Computer Vision, particularly in the realm of medical image analysis. His in-depth research interests include the Swin Transformer, Attention Mechanism-based architecture, and its development for classification problems in the field of health informatics. He is actively involved in research under the Data Science and Intelligent Systems (DSIS) Expertise Group.