RESEARCH ARTICLE OPEN ACCESS

Swin Transformer for Breast Cancer Classification using Histopathology Images

Yash Vishwakarma [1], Akhilesh A. Waoo [2]
Department of CSE, AKS University, Satna - India
Department of CSE, AKS University, Satna - India

ABSTRACT
Cancer of the breast, often known as breast cancer (BC), is one of the most lethal types of disease, and it causes lots of deaths among women around the world. Mammography and ultrasonography are the imaging procedures that are typically used for screening for breast cancer. On the other hand, current imaging technologies are unable to distinguish between distinct subtypes of benign and malignant tumors. In this regard, images obtained through histopathology may provide improved sensitivity toward benign and malignant forms of cancer. Because of the success, they have had in a variety of computer vision tasks, Vision Transformer have recently garnered more attention for medical imaging tasks. The Swin Transformer is based on the idea of limiting the self-attention mechanism to non-overlapping shifted windows and has demonstrated its effectiveness for a range of computer vision tasks. Hence, the goal of this study was to investigate the proficiency of the Swin Transformer model in the binary classification of benign and malignant tumors utilizing a publically available dataset known as BreaKHis, which has 7909 histopathological images obtained at various zoom factors, namely 40X, 100X, 200X, and 400X. The custom Swin Transformer model was trained on the dataset from scratch without using any pre-trained weights, which is typical in most studies involving Transformer, and yet it achieved an average test-case accuracy of 94.05%, this result outperformed all SOTA works. As a result, the custom Swin Transformer model was able to identify BC subtypes based on histopathological images, which can ease the workload of pathologists.

Keywords — swin transformer, histopathology images, binary classification, breast cancer classification, medical imaging

I. INTRODUCTION

Currently, cancer is one of the primary causes of illness and mortality on a global scale. In 2020 alone, it was responsible for the deaths of approximately 10 million individuals. There are projected to be 28.4 million newly diagnosed instances of cancer in the year 2040. The projected number of cancer diagnoses in the year 2040 is a sharp rise of 47% when compared to the year 2020. The American Cancer Society (ACS) reports that lung cancer has been eclipsed by female breast cancer as the most often diagnosed form of cancer. They anticipate 2.3 million new cases of breast cancer in women. There are approximately 6.9% of all cancer-related deaths due to breast cancer, making it the fifth most prominent cause of death worldwide due to cancer (0.68 million in 2020). This places breast cancer behind lung, colorectal, liver, and stomach cancers as the leading cause of cancer-related mortality. Additionally, breast cancer in females accounts for 11.7% of all newly diagnosed malignancies, making it the most prevalent form of the disease [1].

A breast biopsy and subsequent microscopic image analysis are required to obtain a diagnosis of breast cancer. The pathologist can examine the breast tissue's microscopic architecture and the components of the breast tissue by using a sample of breast tissue. The pathologist is then able to categorize these histological images into normal tissue, non-malignant (benign) tissue, and malignant lesions. After that, the information is used for analysis, and a diagnosis is made based on that analysis [2].

The development of breast cancer is not linked to the presence of benign breast lesions, which are anomalies in normal breast tissue. In-situ cancer tissue and invasive cancer tissue are the two types of cancerous tissue. The term "in-situ tissue" refers to the tissue that is found within the ductal-lobular mammary gland. On the other hand, invasive cancer cells were able to spread outside of the ductal and lobular structure of the breast. Before the eye expert examines the biopsied tissue samples, the samples are stained with hematoxylin and eosin (H&E). During the diagnostic technique, whole-slide tissue scans are used to determine the location of the diseased region [3].

In addition to this, the pathologist looks at micrographs of tissue samples at different magnification levels. Multiple image features are analyzed to determine the proper diagnosis [4]. When conducting analysis on images with variable magnification factors, it is necessary to pan, zoom, and focus on each image in addition to performing a comprehensive scan of each image. The identification of breast cancer with this method is likely to produce inaccurate results because it is both time consuming and stressful. Since the beginning of this century, significant advancements have been made in digital imaging. As a direct consequence of these advancements, numerous computer vision and machine learning algorithms developed especially for the purpose of analyzing histopathology images with microscopic resolution have seen light [5] [6] [7] [8] [9].

It is possible that these technologies will make it easier to automate pathological process operations within a diagnostic system. In spite of this, a method of image processing that is both effective and reliable is required for clinical applications.

Unfortunately, the results of standard tactics do not live up to expectations. As a consequence of this, we are still a long way from being able to automatically detect breast cancer based on histology images [10]. Meanwhile, many breakthroughs in Deep Learning techniques have showed significant promise with performance that is state-of-the-art (SOTA) on varieties of tasks. These algorithms have been utilized in a wide number of medical imaging modalities, like histopathology imaging, and have demonstrated enhanced classification, segmentation, and detection capabilities [11]. These methods have proven to be quite useful in medical imaging; nevertheless, they call on a substantial amount of labelled or annotated data, which is currently lacking in this application domain for a number of different reasons [12]. In addition to everything else, annotating a dataset is a particularly time-consuming and expensive process [13].

Vision transformer (ViT) [14] has recently demonstrated its capabilities in image processing tasks by achieving results that are equal to those obtained by models based on convolutional neural networks (CNN) while using significantly fewer computer resources. The self-attention architecture-based Transformer [15] model for natural language processing (NLP) is quickly becoming the industry standard for NLP tasks [16]. The training pace for natural language processing (NLP) tasks can be greatly boosted by applying attention models, also known as transformer. As a result, the performance of neural machine translation applications can be significantly improved. Vision transformer are becoming more prevalent and are starting to show their potential for image processing by being applied to computer vision applications such as image recognition. Rarely does ViT employ convolutional filters, the core of CNNs [17]. Typically, convolutional filters are used for tokenization. Therefore, ViT structurally lacks locality inductive bias compared to CNNs, which require ViT an excessive amount of training data to produce an acceptable visual representation [18]. ViT performs better when pretrained on sufficient data, surpassing a comparable state-ofthe-art CNN with fewer computational resources. To minimize the cost of pre-training, a number of ViTs capable of learning a medium-sized dataset from scratch, such as ImageNet [19], have been proposed. In terms of network architecture, these data-efficient ViTs attempted to amplify the locality inductive bias. Some chose a hierarchical structure, such as CNNs [20], to exploit multiple receptive fields, while others attempted to alter the self-attention mechanism [21]. Swin Transformer [22] are hierarchical vision transformer that make use of shifting windows. They are an enhanced variant of the architecture known as ViT. In order to aid accurate modelling, the idea of self-attention was first introduced within the setting of local windows, and then its computation was performed. Additionally, in order to ensure that the image was evenly partitioned, the windows were placed in a manner that prevented them from overlapping one another.

The window-based method of self-attention has a degree of complexity that may be thought of as linear, and it can be easily scaled. Nevertheless, the value of window-based modelling is restricted as its ability to self-attention has its bounds because it does not allow for connections to be made between other windows. Consequently, a shifted window partitioning strategy was developed. This strategy uses sequential Swin transformer blocks and switches between the various configurations of partitioning. This made it feasible to build cross-window connections while also ensuring that the computation of non-overlapping windows was carried out in the most efficient manner. The shifted window strategy that is used in Swin transformer offers greater efficiency as a result of the restriction of self-attention mechanism to non-overlapping shifted windows. Additionally, this strategy makes it possible to connect windows that are located in different locations. In comparison to that of the ViTs, the performance of the Swin Transformer network was significantly higher. However, learning from scratch on datasets of medium size continues to incur high costs. In addition, learning small-scale datasets from scratch is extremely difficult due to the trade-off between dataset size and performance.

To our knowledge, there is no study that attempts to automate the binary-class classification of breast cancer based on histopathological images utilizing Swin Transformer without using any pre-training techniques. Additionally, we have compared our proposed custom Swin Transformer model with ten distinct pre-trained DL models. Our experimental results demonstrate the robustness of the proposed model for the accurate classification of breast cancer histopathology images. As a result, the primary contributions that come from this research are as follows:

- We propose custom Swin Transformer, a DL model for automated classification of breast cancer histopathology images.
- We trained the existing Swin Transformer with custom parameters and hyperparameters from scratch, without any transfer learning approaches to investigate the model's capability for binary (benign vs. malignant) classification.
- In addition, for the binary-class classification, we utilized the BreakHis dataset in its entirety with regards to the images obtained at 40x, 100x, 200x, and 400x zoom factors.
- In the end, cross-validation and testing were carried out on each of the specific classifications of the distinct zoom factors.
- The proposed custom Swin Transformer model is evaluated in terms of various performance metrics, such as accuracy, F_1-score, sensitivity, and precision. We have also compared our proposed model with some pre-trained DL models and a state-of-the-art model.

The sections of the paper are arranged as follows: In Section II, the material and methodology are described. The results of the experiment are discussed in Section III, and the paper is concluded in Section IV.

II. MATERIAL AND METHODOLOGY

A. BreakHis Dataset

The dataset was split into three sets, with 70% of the data being utilized for training, 10% for validation, and 20% for testing. The original dataset's image intensity values ranged from 0 to 255. The intensities underwent rescaling through the utilization of the pre-processing technique provided by Tensorflow. A resolution of 224 x 224 pixels was utilized for the training of Swin Transformer model. The resolution was kept the same for a fair comparison with all the ten pre-trained and fine-tuned DL models. Both benign and malignant tumors are separated into their own categories within the BreakHis dataset. If a lesion does not meet any of the criteria for cancer, then its histological classification is that it is benign, such as marked cellular atypia, mitosis, basement membrane breakdown, metastasizing, and so on. Benign tumors are typical "innocents," as their growth is modest and confined. Cancer is referred to as a malignant tumor, which is a lesion that has the potential to infiltrate and damage neighboring structures (known as "locally invasive") as well as spread to distant areas (known as "metastasize"), ultimately leading to death. BreakHis is a dataset that contains 7,909 microscopic images of breast tumor tissue. These images were taken from 82 patients and were captured at a variety of magnification factors (40x, 100x, 200x, and 400x). It consists of 2,480 samples that are benign and 5,429 samples that are malignant. These images have a resolution of 700 x 460 pixels and are of RGB type with 8-bit depth in each of the three channels and are provided in PNG format [23]. The photographs were taken by P&D Laboratory in Brazil between January 2014 and December 2014. In the current version of the dataset, samples were obtained through the use of the SOB technique, also called as partial mastectomy or excisional biopsy. This procedure, in contrast to needle biopsy techniques, retrieves a larger tissue sample and is conducted under general anaesthesia in a hospital setting. The number of images with respect to their classes and zoom factors can be seen in table I.

TABLE I
THE NUMBER OF IMAGES AT VARIOUS ZOOM FACTORS FROM BREAKHIS DATASET UNDER BENIGN AND MALIGNANT CLASSES.

Zoom	Benign Samples	Malignant Samples	Total
40x	625	1370	1995
100x	644	1437	2081
200x	623	1390	2013
400x	588	1232	1820
Total	2480	5429	7909

B. Swin Transformer Model

Figure 2 comprehensively summarizes the Swin Transformer model's architectural components. From its initial 700x460 pixels, the image was resized to just 224 x 224 pixels. Due to the limitations of our system's processing power and memory limitations, this was the best choice. In addition, the original patch size of 4 × 4 is replaced with a smaller starting patch size of 2 x 2, and the input RGB picture with input dimensions of H x W x 3 is broken into small patches having sizes equal to 2 x 2. Therefore, each patch is of the size 2 x 2 x 3 = 12. A linear embedding layer is then applied on top of this raw 12-sized feature tensor

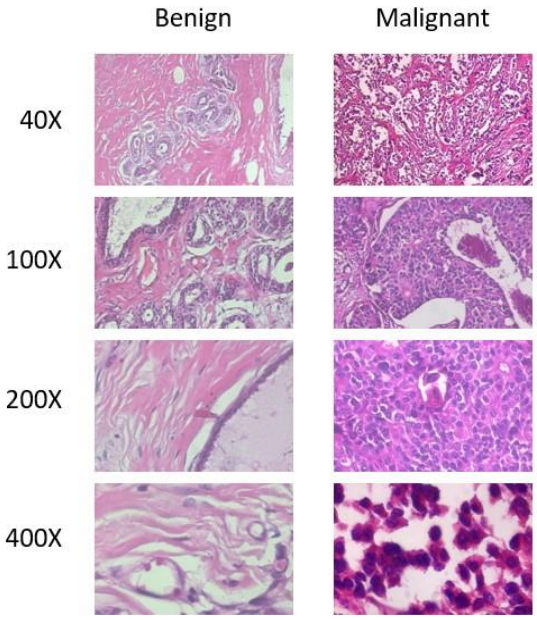

Fig. 1 Sample of images from the dataset at various zoom factors.

to allow for projection onto an arbitrary feature dimension denoted as C. Keeping the token count constant at $\frac{H}{4} \times \frac{W}{4}$, these patch linear embeddings are subjected to various Swin transformer blocks that have a modified self-attention. Stage 1 of the Swin transformer architecture includes the linear embedding layer and Swin transformer blocks. To simplify the representation of data in a hierarchical structure and reduce the number of patches required, the Swin Transformer design introduces patch merging layers at Stage 2. In Stage 2, the patch merging layer adds a linear layer to the aggregated features from all 4C dimensions, which are the sum of the features from each pair of adjacent 2x2 patches. After this, the linear layer's output depth is updated as 2C, and the total number of patches is reduced by 2 x 2 = 4. The number of Stage 2 output patches is kept constant at $\frac{H}{8} \times \frac{W}{8}$, and feature transformation is accomplished with the help of Swin transformer blocks. Stages 3 and 4 repeat this procedure twice more, yielding an output resolution of either $\frac{H}{16} \times \frac{W}{16}$ or $\frac{H}{32} \times \frac{W}{32}$, depending on the number of stages performed.

The combined effect of these steps yields a deep hierarchical representation with feature map sizes common to popular CNN models like VGGNet [24] and ResNet

[25]. This means the architecture has a good shot at replacing the previous approaches' use of backbone networks in the numerous vision activities that are currently being completed.

1) ***Swin Transformer Block:*** The standard multi-head self-attention (MSA) module in a Transformer block is swapped out for a module based on shifted windows during the construction of a Swin Transformer. The other layers of the Transformer block are left unchanged throughout this process.

conventionally partitioned windows by a distance equivalent to $\frac{M}{4}, \frac{M}{4}$ pixels. When using the strategy known as shifting window partitioning, the consecutive blocks are computed as

$$z^{l} = W - MSA(LN(z^{l-1})) + z^{l-1} \quad (2)$$

$$z^{l} = MLP(LN(z^{l})) + z^{l} \quad (3)$$

$$z^{l+1} = SW - MSA(LN(z^{l})) + z^{l} \quad (4)$$

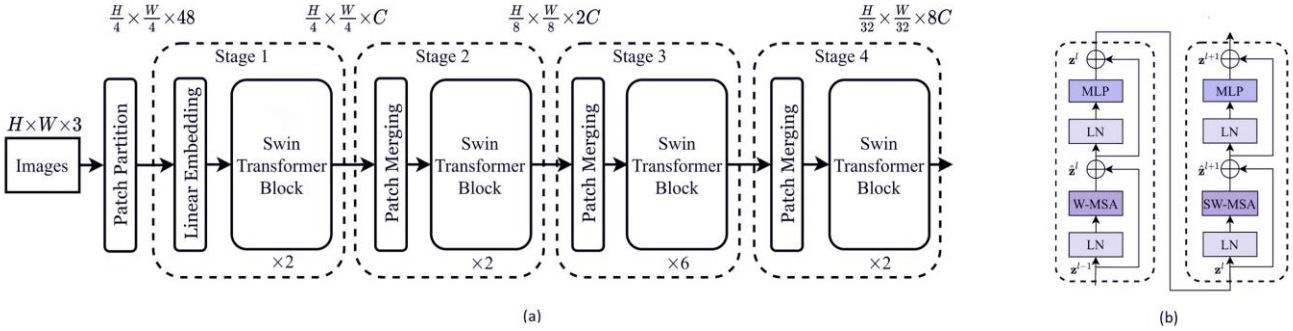

Fig. 2 (a) Architecture of Swin Transformer model [22], (b) Two consecutive Blocks

A shifted window-based MSA module is the first component of a Swin Transformer block. Following this is a two-layered Multilayer perceptron (MLP) with GELU [26] nonlinearity embedded in the middle. A Layer Normalization (LN) layer before each MSA module and MLP, and a residual connection is added after each module. The Gaussian Error Linear Unit, or GELU, is an activation function which produces smoother curve values when compared to RELU and has the following formula:

$$GELU(x) = x\Phi(x) \quad (1)$$

2) ***Self Attention using Shifted Windows:*** The window-based self-attention module's modelling capacity is restricted due to the absence of inter-window connections. The utilization of a shifted window partitioning technique is employed to ensure the efficient computation of non-overlapping windows. The aforementioned approach alternates between two distinct partitioning configurations within consecutive Swin Transformer blocks. This will enable us to establish cross-window connections. The initial module is accountable for implementing a conventional technique of window partitioning, commencing from the pixel located in the upper-left corner. Subsequently, the 8 × 8 feature map undergoes uniform partitioning into 4 x 4 windows of dimensions 4 x 4, where M is equal to 4. Subsequently, the next module employs a distinct windowing configuration in contrast to its preceding layer. The aforementioned task is achieved through the displacement of the windows from the

$$z^{l+1} = MLP(LN(z^{l+1})) + z^{l+1} \quad (5)$$

where "z^{l}" and "z^{l+1}" denote the output features of the (S)WMSA module and the MLP module for block l, respectively while "W-MSA" and "SW-MSA" denote window based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

The self-attention is computed by adding a relative position bias (denoted by "B") to every head.

$$B \in \mathbb{R}^{M^2 \times M^2} \quad (6)$$

$$Attention(Q, K, V) = SoftMax(QK^T\sqrt{d} + B)V \quad (7)$$

where "Q", "K", "V" $\in R^{M^2 \times d}$ are the query, key and value matrices; "d" is the query/key dimension, and "M^2" is the number of patches in a window.

The Sigmoid activation function was used to classify the inputs into two distinct classes, and it has the following formula:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (8)$$

C. Model Validation and Testing

The dataset was split into three sets, with 70% of the data being utilized for training, 10% for validation, and 20% for testing. The image size was kept constant at 224 x 224

pixels for all training, validation, and testing purposes. The original dataset's image intensity values ranged from 0 to 255. The intensities underwent rescaling through the utilization of the pre-processing technique provided by Tensorflow. After training the model the validation set of images was analyzed to ensure that the model did not contain any instances of overfitting.

D. *Performance Metrics*

In addition to accuracy, other performance metrics, including weighted-averaged versions of precision, sensitivity, and F1-score, were utilized because the number of samples in both the benign and malignant, considering zoom factors, were imbalanced. The number of incorrect classifications that lie above the off-diagonal in the confusion matrix was counted as false positives, while the number of incorrect classifications that fell below the off-diagonal was counted as false negatives. The true negatives represent the number of items that have been accurately categorized for other classes besides that particular class. "TP", "TN", "FP", and "FN" denote "true positives," "true negatives," "false positives," and "false negatives," respectively. These are the foundation for the mathematical formulae that are used to calculate the performance measures described above. The mathematical equations used for the calculation of performance metrics are given below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (12)$$

III. EXPERIMENTAL RESULTS

A. *Experimental Settings and Network training*

The Swin Transformer model was trained for 100 epochs with Adam optimizer, a learning rate of 0.001, a batch size of 128, a weight decay factor of 0.0001, patch size of (2, 2), dropout rate of 0.05, attention heads equal to 8, embedding dimension of 256, MLP layer size of 256, attention window of size equal to 4, size of shifting window equal to 1, and image size equal to 224 x 224 pixels. The dataset was split into three sets, with 70% of the data being utilized for training, 10% for validation, and 20% for testing.

B. *Classification Results*

The custom Swin Transformer performed very well for binary classification (benign vs malignant) among all the zoom factors. The model had the best performance on 400x magnification in which it attained a test accuracy of 95.24%, sensitivity of 96.33%, precision of 94.06%, and F1-score of 95.18%. The performance metrics for all the zoom factors are given in the table II. We have also included confusion matrices for each magnification level in figures 3, 4, 5, and 6.

TABLE II
RESULTS FOR THE BINARY CLASSIFICATION ZOOM FACTOR WISE.

Zoom	Accuracy	Sensitivity	Precision	F1-score
40x	92.50	93.52	91.33	92.41
100x	94.55	95.37	93.64	94.50
200x	93.90	94.86	92.83	93.83
400x	95.24	96.33	94.06	95.18

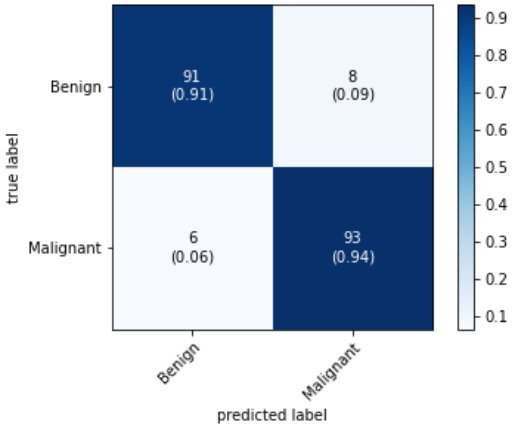

Fig. 3 Confusion Matrix for 40X.

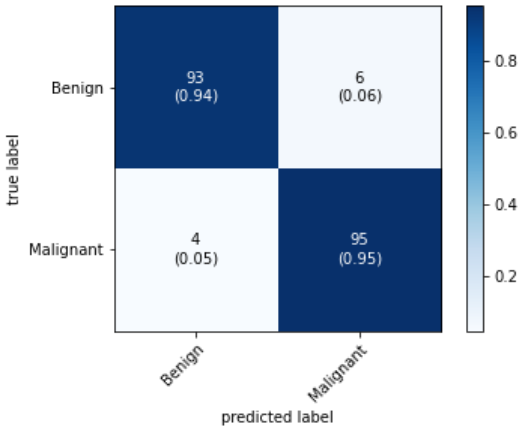

Fig. 4 Confusion Matrix for 100X.

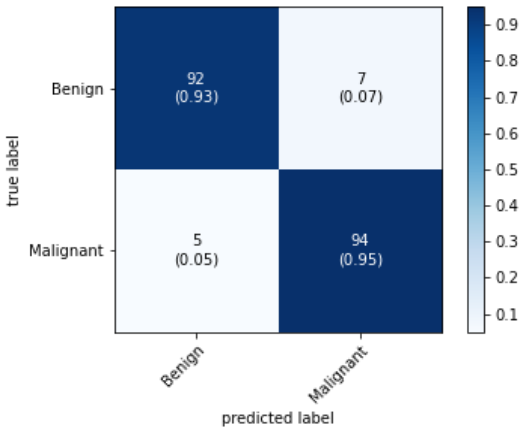

Fig. 5 Confusion Matrix for 200X.

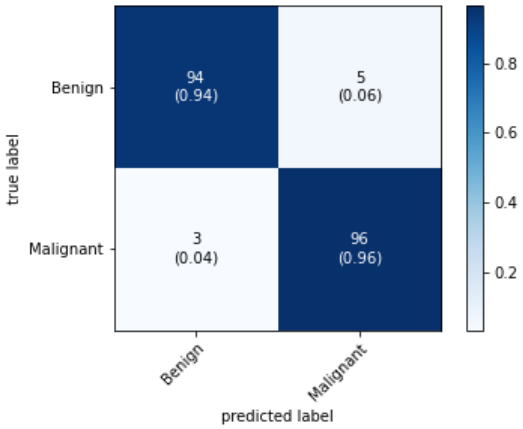

Fig. 6 Confusion Matrix for 400X.

C. Comparison with common pre-trained networks

The proposed custom Swin Transformer model was compared with 10 distinct pre-trained fine-tuned DL models. These pre-trained models VGG-16, VGG-19, ResNet-50, ResNet-101, ResNet-152, InceptionV3 [27], Xception [28], MobileNetV2 [29], DenseNet121 [30], Vision Transformer (ViT) are pre-trained on ImageNet dataset which consists of more than 1000 classes. The test accuracies of these models were averaged among all the zoom factors and are shown in table III. In comparison to the other pre-trained models, the proposed custom Swin Transformer model exhibited significant performance and superior convergence with an increasing number of epochs, and it also outperformed all the pre-trained models by a large margin. Learning behaviours of our model and pre-trained networks are shown in figures 7 and 8.

TABLE III
AVERAGE TEST CASE ACCURACY COMPARISON WITH PRE-TRAINED NETWORKS.

Architecture	Accuracy (%)
VGG-16	85.49
VGG-19	85.01
ResNet-50	68.50
ResNet-101	70.49
ResNet-152	71.49
InceptionV3	71.49
Xception	64.50
MobileNetV2	66.77
DenseNet-121	73.99
Vision Transformer	93.09
This Work	**94.05**

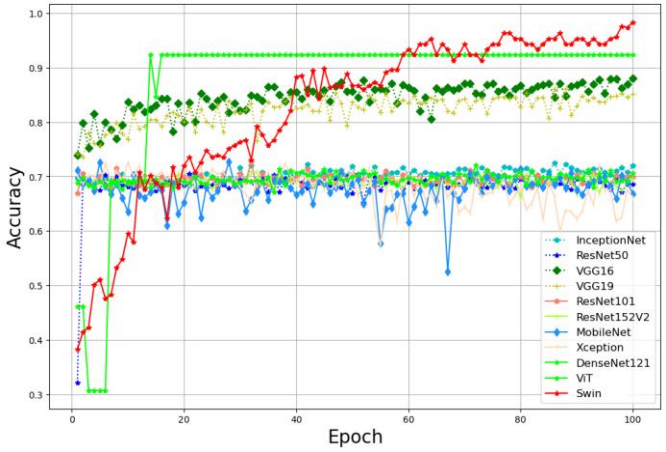

Fig. 7 Validation accuracy of custom Swin Transformer and Pre-trained networks over the epochs.

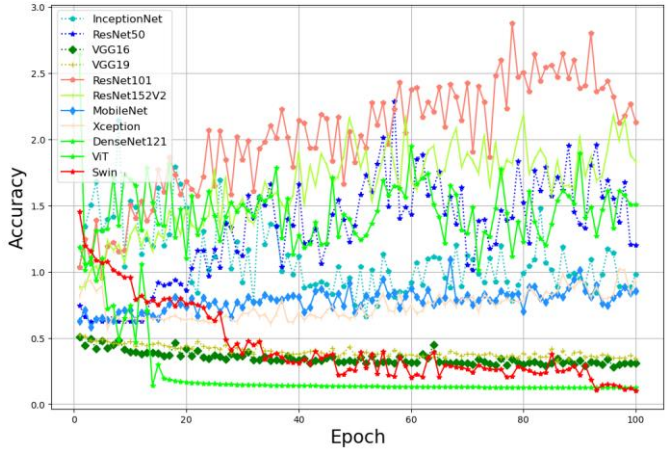

Fig. 8 Validation loss of custom Swin Transformer and Pre-trained networks over the epochs.

D. Comparison with prior SOTA

The proposed custom Swin Transformer model was also compared with existing SOTA models for binary classification of BreakHis in table IV. And our model performed better than all the SOTA models at 40X, 100X, 200X, and 400X zoom factors.

TABLE IV
PRIOR SOTA ACCURACY COMPARISON FOR BINARY CLASSIFICATION OF BREAKHIS.

Ref	Method	40X	100X	200X	400X
[31]	LPQ, SVM	91.10	90.70	86.20	84.30
[32]	Boltzmann Machine	88.70	85.30	88.60	88.40
[33]	CNN	89.52	89.06	88.84	87.67
	This Work	**92.50**	**94.55**	**93.90**	**95.24**

IV. CONCLUSION

This study proposes a custom Swin Transformer model for the binary classification of benign and malignant classes zoom-wise from the BreaKHis dataset. Taking all the magnifications into account, our model performed better than all SOTA models for breast cancer histopathology binary classification with an average test accuracy of 94.05%. Because of this, the custom Swin Transformer model has the potential to be utilized for computer-assisted diagnosis of benign and malignant conditions, which will result in accurate diagnoses and will assist in reducing the burden of pathologists.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.

[3] S. Chattoraj, S. Pratiher, R. Mukherjee, S. Ghosh, A. Chakraborty, D. Hazra, and S. Pratiher, "Efficacy of deep convolutional neural network features on histological manifold for robust breast carcinoma detection," in *SPIE/COS Photonics Asia*, 2018.

[4] M. Peikari, M. J. Gangeh, J. Zubovits, G. Clarke, and A. L. Martel, "Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 307–315, 2015.

[5] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[6] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, p. 221, 2017.

[8] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[9] S. Chattoraj and K. Vishwakarma, "Classification of histopathological breast cancer images using iterative vmd aided zernike moments & textural signatures," *ArXiv*, vol. abs/1801.04880, 2018.

[10] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps*, pp. 323–350, 2018.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sanchez,′ "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[12] F. Altaf, S. M. Islam, N. Akhtar, and N. K. Janjua, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99540–99572, 2019.

[13] D. Nawn, S. Pratiher, S. Chattoraj, D. Chakraborty, M. Pal, R. R. Paul, S. Dutta, and J. Chatterjee, "Multifractal alterations in oral sub-epithelial connective tissue during progression of pre-cancer and cancer," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 152–162, 2020.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[17] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[18] B. Neyshabur, "Towards learning convolutions from scratch," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8078–8088, 2020.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11936–11945.

[21] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," *arXiv preprint arXiv:2104.05707*, 2021.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[23] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] J. A. Badejo, E. Adetiba, A. Akinrinmade, and M. B. Akanle, "Medical image classification with hand-designed or machine-designed texture descriptors: a performance evaluation," in *international conference on bioinformatics and biomedical engineering*. Springer, 2018, pp. 266– 275.

[32] A.-A. Nahid, A. Mikaelian, and Y. Kong, "Histopathological breastimage classification with restricted boltzmann machine along with backpropagation," *Biomedical Research*, vol. 29, no. 10, pp. 2068–2077, 2018.

[33] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 578–581, 2018.