Magnification Invariant Medical Image Analysis: A Comparison of Convolutional Networks, Vision Transformers, and Token Mixers

Pranav Jeevan [1] **Nikhil Cherian Kurian** [1] **Amit Sethi** [1]

Abstract

Convolution Neural Networks (CNNs) are widely used in medical image analysis, but their performance degrade when the magnification of testing images differ from the training images. The inability of CNNs to generalize across magnification scales can result in sub-optimal performance on external datasets. This study aims to evaluate the robustness of various deep learning architectures in the analysis of breast cancer histopathological images with varying magnification scales at training and testing stages. Here we explore and compare the performance of multiple deep learning architectures, including CNN-based ResNet and MobileNet, self-attention-based Vision Transformers and Swin Transformers, and token-mixing models, such as FNet, ConvMixer, MLP-Mixer, and WaveMix. The experiments are conducted using the BreakHis (Spanhol et al., 2015) dataset, which contains breast cancer histopathological images at varying magnification levels. We show that performance of WaveMix is invariant to the magnification of training and testing data and can provide stable and good classification accuracy. These evaluations are critical in identifying deep learning architectures that can robustly handle changes in magnification scale, ensuring that scale changes across anatomical structures do not disturb the inference results.

1. Introduction

Computer aided medical image analysis has become a critical component in the diagnosis and treatment of various diseases (Chakraborty & Mali, 2023; Duncan & Ayache, 2000). Deep learning models, such as Convolution neural networks (CNNs), have shown exceptional performance in analyzing medical images, including magnetic resonance imaging (MRI), computed tomography (CT), and histology images (Chan et al., 2020). However, the performance of these models can be affected by several factors, including variations in image quality, lighting conditions, and magnification scales. In particular, changes in magnification scales between training and testing datasets can significantly impact the accuracy and robustness of deep learning models in medical image analysis (Gupta & Bhavsar, 2017).

CNNs are cited to be the most commonly used deep learning architecture for medical image analysis (Li et al., 2014). However CNN, can struggle when it comes to handling medical images with anatomical features at varying magnification scales. In general, training a CNN on images at a specific magnification scale may result in good performance on that scale, but this performance may not generalize well to other magnification scales (Alkassar et al., 2021). This is a significant limitation when analysing medical imaging modalities like histology images where slight to moderate changes in magnification variability is common. The inability of CNN to generalize across magnification scales leads to sub-optimal inference performance on external datasets (Gupta & Bhavsar, 2017). Though, augmenting input images with perturbations in scales can slightly improve performance of CNNs, it is also important to explore or develop more robust deep learning architectures that can generate features that are inherently invariant to the changes in scale of input images. Such architectures should be designed to capture the important features in the images, regardless of the shift in the magnification scale, in order to provide robust performance for medical image analysis in a clinical settings.

In this study, we evaluate the robustness of multiple popular deep learning architectures including CNN based architectures such as ResNet (He et al., 2016) and MobileNet (Howard et al., 2017), Self-attention based architectures such as Vision Transformers (VIT) (Dosovitskiy et al., 2021) and Swin Transformers (Liu et al., 2021), and token mixing models such as Fourier-Net (FNet) (Lee-Thorp et al., 2021), ConvMixer (Trockman & Kolter, 2022), Multi-Layer Perceptron-Mixer (MLP-Mixer) (Tolstikhin et al., 2021), and WaveMix (Jeevan et al., 2022). Our aim is to compare the performance of these deep learning models when the

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India. Correspondence to: Pranav Jeevan <194070025@iitb.ac.in>.

magnification of the test data differs from the training data. The BreakHis (Spanhol et al., 2015) dataset , which includes breast cancer histopathological images at varying magnification levels, is utilized for our experiments. The empirical performance differences between the deep learning models will be used to determine the most robust architecture for histopathological image analysis.

2. Experiments

2.1. Dataset

We utilize the BreakHis (Spanhol et al., 2015) dataset, which is a well-known public dataset in the field of digital breast histopathology for our experiments. It has been widely used in the development and evaluation of computer-aided diagnosis (CAD) systems for breast cancer diagnosis. It provides a challenging benchmark for the development of CAD systems due to the inherent large variations in tissue appearances.

The dataset consist of 7,909 microscopy images of breast tissue biopsy specimens from 82 patients diagnosed with either benign or malignant breast tumors. The images are collected from four different institutions and are of four different magnifications scales - 40X, 100X, 200X and 400X.

In addition to the malignancy information of each image, the dataset is further annotated with information like the patient's age, the sub-type of malignancy and the type of biopsy. The dataset is slightly imbalanced in terms of the distribution of benign and malignant cases and the distribution of different magnifications. In the dataset there are 5,429 malignant cases whereas benign cases are only about 2,480.

As the BreakHis (Spanhol et al., 2015) dataset contains multiple images at different magnification levels, the dataset serves as a challenging and representative testbed for evaluating the robustness of deep learning architectures across the different magnification levels or scales. These evaluations will be carried out by training some of the recently reported deep learning architecture on one magnification level of the BreakHis (Spanhol et al., 2015) dataset and testing these trained models across multiple held-out magnification levels. Observing the average test accuracy on the different magnification levels can hence reveal the robustness of deep learning architectures to varying image magnification at inference..

2.2. Models

For CNN based models, we compared performance using ResNet-18, ResNet-34 and ResNet-50 from the ResNet family (He et al., 2016), and MobileNetV3-small-0.50, MobileNetV3-small-0.75 and MobileNetV3-small-100 from MobileNet family of models. We used ViT-Tiny, ViT-Small and ViT-Base (all using patch size of 16, see (Dosovitskiy et al., 2021)) along with Swin-Tiny and Swin-Base (all using patch size of 4 and window size of 7, see (Liu et al., 2021)) for the experiments.

2.2.1. Token-Mixers

Token-mixers are the family of models which uses an architecture similar to MetaFormer (Yu et al., 2022) as its fundamental block as shown in Figure 1. Transformer models can be considered as token-mixing model which uses self-attention for token-mixing. Other token-mixers use Fourier transforms (FNet) (Lee-Thorp et al., 2021), Wavelet transforms (WaveMix) (Jeevan et al., 2022), spatial-MLP (MLP-Mixer) (Tolstikhin et al., 2021) or depth-wise convolutions (ConvMixer) (Trockman & Kolter, 2022) for token-mixing. Token-mixing models have been shown to be more efficient in terms of parameters and computation compared to attention-based transformers (Yu et al., 2022).

FNet (Lee-Thorp et al., 2021) was actually designed for natural language processing (NLP) tasks and was designed to handle 1D inputs sequences. We have used the 2D-FNet, i.e., a modified FNet that used a 2D Fourier transform for spacial token-mixing instead of a 1D Fourier transform used in FNet. The 2-D FNet can process images in the 2D form without the need to unroll it into sequence of patches or pixels as done in transformer and FNet. We experimented by varying the embedding dimension and number of layers to get the best model.

WaveMix (Jeevan et al., 2022) uses 2D-Discrete Wavelet transform (2D-DWT) for token-mixing. We experimented by varying the embedding dimension, number of layers and number of levels of 2D-DWT used in WaveMix to get the model which gives highest validation accuracy in the dataset.

ConvMixer (Trockman & Kolter, 2022) uses depth-wise convolution for spacial token-mixing and point-wise convolutions for channel toke-mixing. We used ConvMixer-1536/20, ConvMixer-768/32, and ConvMixer-1024/20 available in Timm model library for our experiments.

MLP-Mixer (Tolstikhin et al., 2021) uses spacial MLP and channel MLP to mix tokens. We used MLP-Mixer-Small (patch size of 16) and MLP-Mixer-Base (patch size of 16) in our experiments.

2.3. Implementation details

The dataset was divided into train, validation and test sets in the ratio 7:1:2 for each of the magnification. Due to limited computational resources, the maximum number of training epochs was set to 300. All experiments were done with a single 80 GB Nvidia A100 GPU. *No pre-trained weights*

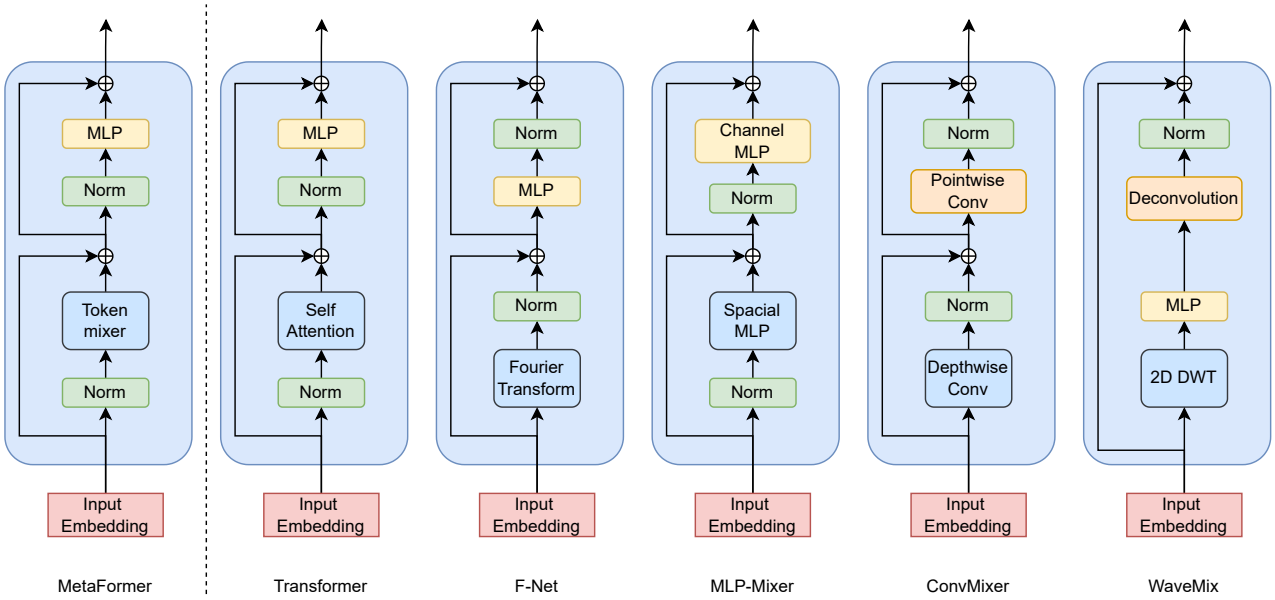

Figure 1. Architectures of various token-mixers along with the general MetaFormer block. The token-mixing operation in different models is performed by different operations, such as spacial MLP, depth-wise convolution, self-attention, Fourier and Wavelet transforms

was used for any of the models. We used the ResNet, MobileNet, Vision transformer, Swin transformer, ConvMixer and MLP-Mixer available in Timm (PyTorch Image Models) library (Wightman, 2019)[1] Since WaveMix and FNet was unavailable in Timm library, it was implemented from original paper. The Timm training script (Wightman, 2019) with default hyper-parameter values was used to train all the models. Cross-entropy loss was used for training. We used automatic mixed precision in PyTorch during training to optimize speed and memory consumption.

The images were resized to 672×448 for the experiments. Transformer-based models and MLP-Mixer required the images to be resized to certain specific sizes like 224×224 or 384×384. We trained models of varying sizes belonging to the same architecture on the training set and evaluated it on validation set to find the model size that gives the best performance on the Breakhis (Spanhol et al., 2015) dataset. The model size with highest average validation performance over all magnifications was used for evaluation using test set.

The maximum batch-size was set to 128. For larger models, we reduced the batch-size so that it can fit in the GPU. Top-1 accuracy on the test set of the best of three runs with random initialization is reported as a generalization metric based on prevailing protocols (Hassani et al., 2021).

[1]available at http://github.com/rwightman/pytorch-image-models/

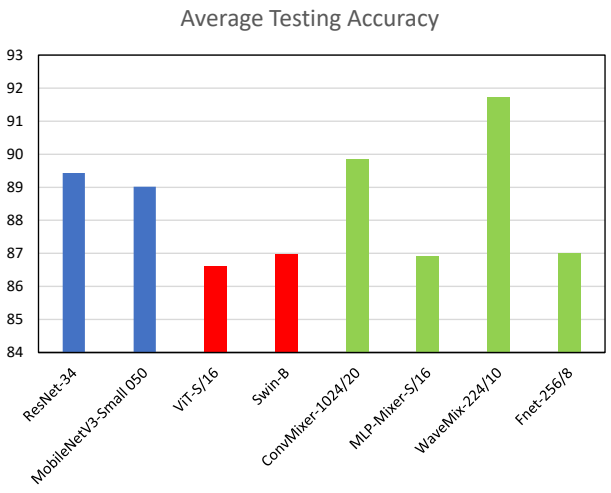

Figure 2. Average of all test accuracy reported for various training magnifications for each of the models compared

3. Results and Discussions

The Inter-magnification classification performance of all the best performing model variants of CNN, transformer and token-mixer models are shown in Table 1. We can see that WaveMix performs better than all the other models in maintaining high performance across different testing magnifications. Only ConvMixer, another token-mixer could

Table 1. Results of Inter-magnification classification performance of all CNN, transformers and token-mixers on Breakhis (Spanhol et al., 2015) dataset. Accuracy on test set is reported.

CNNs

ResNet-34

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	94.74	92.81	81.89	84.11	88.38
100X	88.72	95.20	90.32	90.69	91.23
200X	86.97	89.21	95.53	93.43	91.28
400X	78.20	85.61	87.10	96.44	86.84

MobileNetV3-Small 050

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	92.48	91.13	84.62	82.19	87.60
100X	87.47	89.69	88.59	89.04	88.70
200X	86.97	89.21	94.54	90.96	90.42
400X	85.71	86.81	90.07	94.79	89.35

Transformers

ViT-S/16

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	89.72	86.33	85.11	69.04	82.55
100X	86.72	88.73	87.84	89.86	88.29
200X	86.47	88.49	87.35	88.49	87.70
400X	86.22	87.29	87.59	90.69	87.95

Swin-B

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	91.48	87.05	75.43	70.68	81.16
100X	88.22	88.49	90.57	86.85	88.53
200X	85.97	89.21	92.06	88.22	88.86
400X	87.97	88.01	89.83	91.78	89.40

Token-Mixers

ConvMixer-1024/20

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	96.49	88.49	81.14	81.92	87.01
100X	89.22	96.40	90.07	85.75	90.36
200X	87.47	91.61	96.28	92.33	**91.92**
400X	85.46	88.73	90.57	95.62	90.09

MLP-Mixer-S/16

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	91.98	80.58	78.16	81.10	82.95
100X	86.72	88.73	87.84	89.86	88.29
200X	88.47	88.49	94.29	91.78	90.76
400X	83.46	86.57	84.86	87.67	85.64

WaveMix-224/10

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	95.99	93.77	87.10	90.68	**91.88**
100X	89.97	94.72	92.31	89.86	**91.72**
200X	87.97	89.69	94.79	93.70	91.54
400X	89.31	88.49	91.47	97.69	**91.74**

F-Net-256/8

Training Magnification	Testing Magnification				Average testing performance over all magnifications
	40X	100X	200X	400X	
40X	94.50	85.10	83.90	84.90	87.10
100X	88.70	89.00	84.70	83.40	87.50
200X	86.70	87.10	89.30	88.50	87.90
400X	84.70	82.50	86.40	87.90	85.40

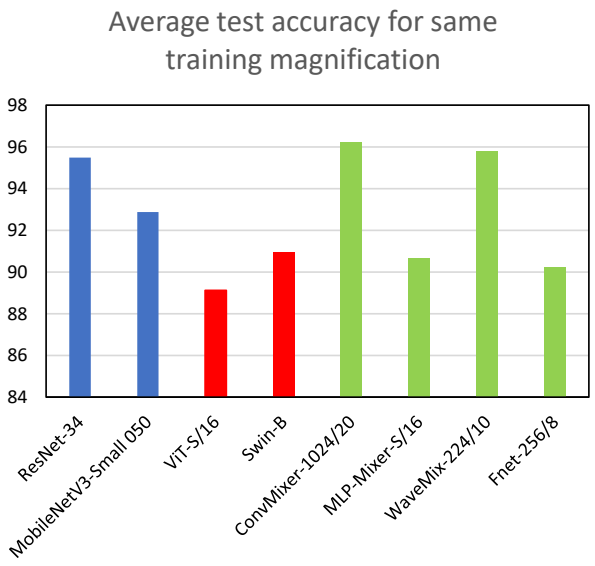

Figure 3. Average of test accuracy when training and testing was done on same magnification for each model is compared

perform better than WaveMix in one magnification (200×). We also observe that the accuracy of WaveMix is the most stable, never falling bellow 87%. Other models that perform well, such as, ConvMixer and ResNet-34, suffers from unstable performance with accuracy falling to 81% and 78% respectively. The better performance of WaveMix is due to the ability of 2D wavelet transform to efficiently mix token information and the subsequent use of deconvolution layers which also aids in rapid expansion of receptive field after each wavelet block.

We also see from Figure 2 that WaveMix performs the best amoung all models when we take the overall average of all the average testing accuracy over all magnifications. We observe that the performance of token-mixers (green) like MLP-Mixer and FNet is comparable to that of transformer based models (red). CNN-based models (blue) perform better than transformer models.

Figure 3 show the average of test accuracy when training and testing was done on same magnifications. We observe that ConvMixer performs better than WaveMix when train and test magnifications are same. Even ResNet-34 is performing almost on par with WaveMix and ConvMixer. This shows that even though other models perform well when magnification of training and test data are same, they cannot translate that performance when magnification of training and testing set differs from each other. WaveMix is mostly invariant to this change of magnification between train and test data and is able to provide consistent performance compared to other CNN, transformer and token-mixing models.

FNet consumed largest GPU RAM (4-8× more) compared to other architectures. CNN-based models perform much better than transformer model-based models in Breakhis classification. There is a significant drop in performance when the transformer-based models are trained on 40× magnification and tested. Similar drop in accuracy for 40× magnification training was observed for MLP-Mixer.

4. Conclusions

In conclusion, our study evaluated the robustness of various deep learning models for histopathological image analysis under different testing magnifications. We compared ResNet, MobileNet, Vision Transformers, Swin Transformers, Fourier-Net, ConvMixer, MLP-Mixer, and WaveMix using the BreakHis (Spanhol et al., 2015) dataset. Our experiments demonstrated that the WaveMix architecture, which intrinsically incorporates multi-resolution features, is the most robust model to changes in inference magnification. We observed a stable accuracy of at least 87% across all test scenarios. These findings highlight the importance of implementing a robust architecture, such as WaveMix, not only for histopathological image analysis but also for medical image analysis in general. This would help to ensure that anatomical features of diverse scales do not influence the accuracy of deep learning-based systems, thereby improving the reliability of diagnostic inference in clinical practice.

References

Alkassar, S., Jebur, B. A., Abdullah, M. A., Al-Khalidy, J. H., and Chambers, J. A. Going deeper: magnification-invariant approach for breast cancer classification using histopathological images. *IET Computer Vision*, 15(2): 151–164, 2021.

Chakraborty, S. and Mali, K. An overview of biomedical image analysis from the deep learning perspective. *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*, pp. 43–59, 2023.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. Deep learning in medical image analysis. *Deep Learning in Medical Image Analysis: Challenges and Applications*, pp. 3–21, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Duncan, J. S. and Ayache, N. Medical image analysis: Progress over two decades and the challenges ahead.

IEEE transactions on pattern analysis and machine intelligence, 22(1):85–106, 2000.

Gupta, V. and Bhavsar, A. Breast cancer histopathological image classification: is magnification important? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 17–24, 2017.

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. Escaping the big data paradigm with compact transformers, 2021.

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

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. April 2017.

Jeevan, P., Viswanathan, K., S, A. A., and Sethi, A. Wavemix: A resource-efficient neural network for image analysis, 2022. URL https://arxiv.org/abs/2205.14375.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee-Thorp, J., Ainslie, J., Eckstein, I., and Ontanon, S. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pp. 844–848. IEEE, 2014.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Trockman, A. and Kolter, J. Z. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10819–10829, 2022.