Early Cancer Detection Using Multi-Scale CNN and Transformer-Based Deep Learning Approaches

Chengdu University of Technology, CDUT Sino-British Collaborative Education

Xinyu Luo(Carson) | 202118020329 | Supervisor: Dr. Grace U. Nneji

Abstract

This study proposes a hybrid deep learning framework for early breast cancer detection based on histopathological image analysis. The model integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), leveraging EfficientNetV2 for efficient local feature extraction and ViTs enhanced with Shifted Patch Tokenization (SPT) for global context modeling. The framework was trained and evaluated on the BreakHis dataset comprising over 7,900 images at multiple magnification levels, and further validated on the BACH dataset of 400 high-resolution microscopy images. On the BreakHis dataset, the model achieved a test accuracy of 92.1%, an AUC-ROC of 0.9715, and a sensitivity of 94.6%, while on the BACH dataset, it attained an accuracy of 86.7%, AUC-ROC of 0.8700, and sensitivity of 92.5%. These results demonstrate the model's strong classification performance and generalization capability across different datasets. The proposed approach shows considerable potential for clinical applications that require reliable, high-sensitivity diagnostic support in digital pathology.

Methodology

Dataset 1: BreakHis

- 1,995 benign, 2,081 malignant images
- 70/10/20 split (train/val/test) with class balancing via random oversampling

Split Strategy

Custom stratified split:

- BreakHis: 70% train, 10% val, 20% test
- BACH: 80% train, 20% test

Dataset 2： BACH

- 100 benign, 200 malignant (after re-labeling)
- 80/20 split with oversampling to balance classes

Preprocessing

- Resize all images to 160×160
- Normalize to [0,1]
- Apply real-time data augmentation:
- Random rotation (±90°), zoom (0.8–1.2×), translation, horizontal/vertical flip

This study proposes a hybrid deep learning model combining Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to classify histopathological images of breast cancer.
The model is trained on the BreakHis dataset (100x magnification), using a dual-branch architecture: The **EfficientNetV2** [1] branch extracts local features through compound convolutional scaling.

The **ViT branch** utilizes **Shifted Patch Tokenization (SPT) [2]** to enhance local spatial representation and **Learned-Scale Attention (LSA)** to adjust attention scores adaptively.

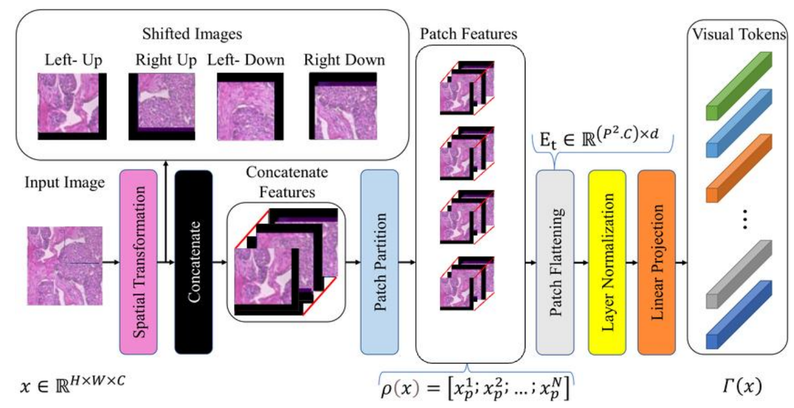


Figure 1: Shifted Patch Tokenization generates 5 shifted views to improve spatial diversity

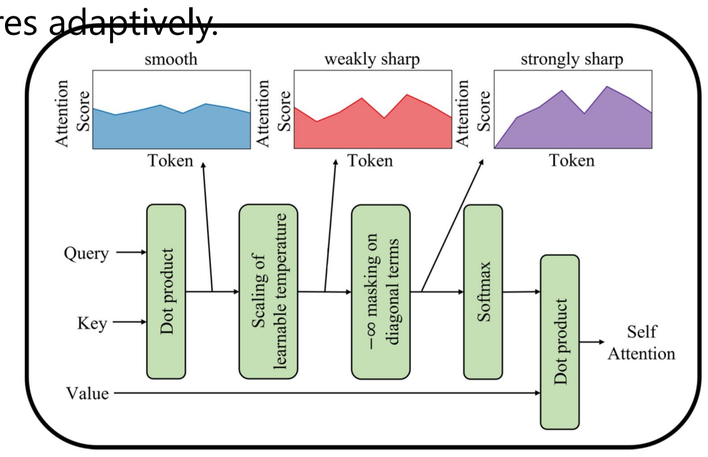


Figure 2: Learned-Scale Attention dynamically adjusts attention scores across heads.

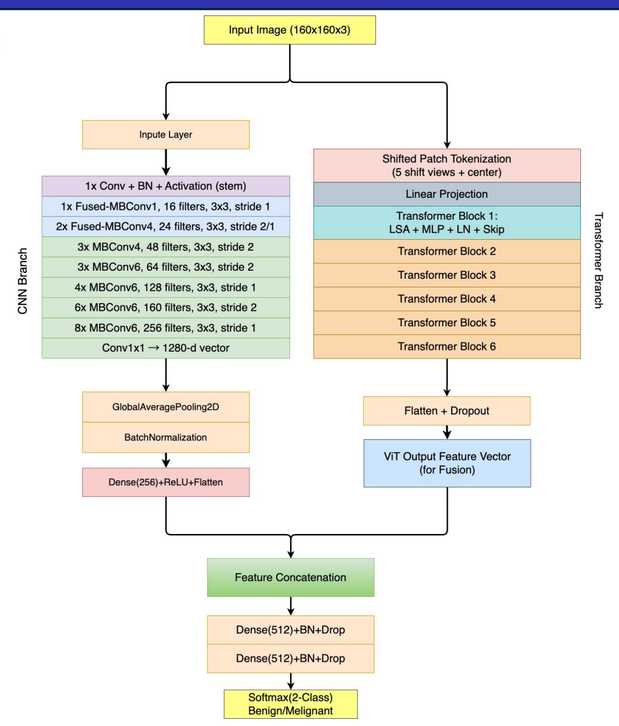
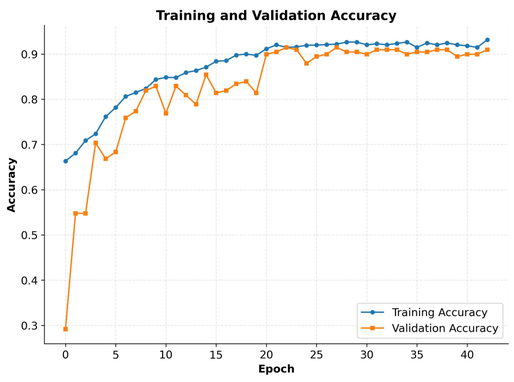


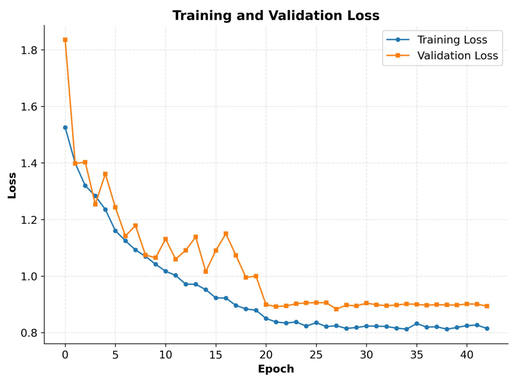
Figure 3: Hybrid architecture combining EfficientNetV2 and ViT for breast cancer classification.

Results

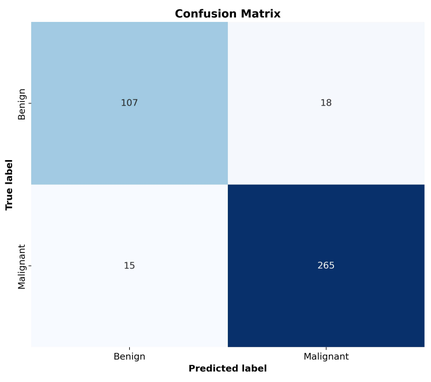
BreakHis Dataset



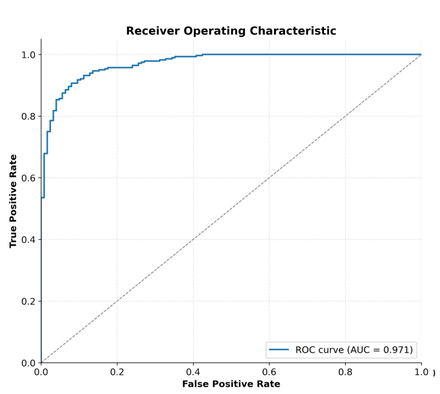
train_acc=0.9317
val_acc=0.9095



train_loss=0.8145
val_loss=0.8934



TN=107, TP=265, FP=18, FN=15



AUC=0.971

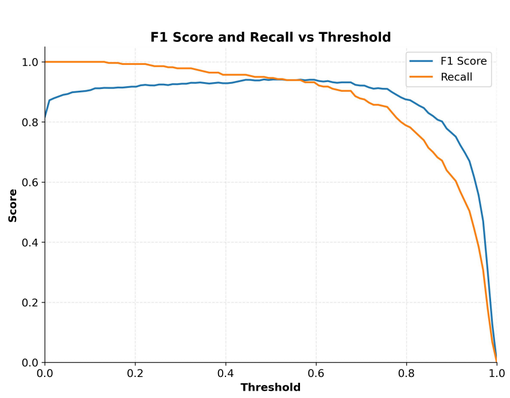
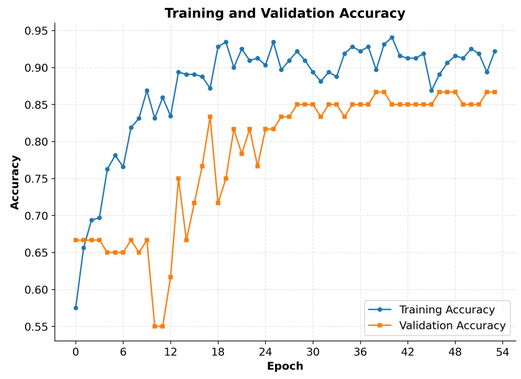
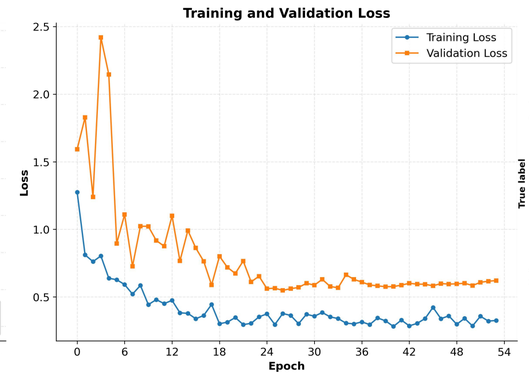


Figure 4: Training Resulets Summary of BreakHis Dataset

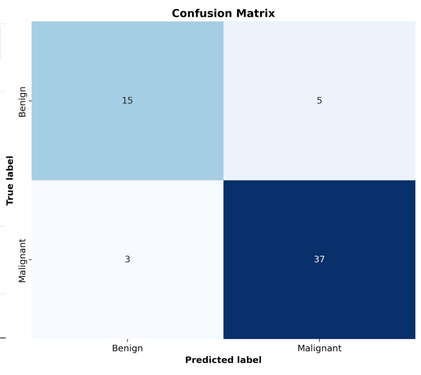
BACH Dataset



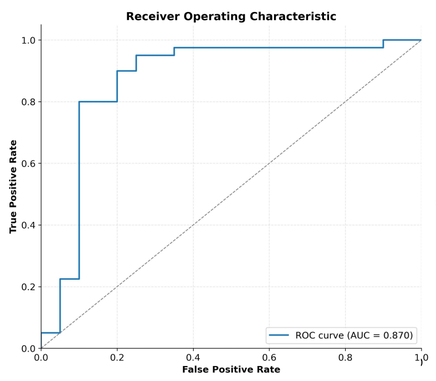
train_acc=0.9219
val_acc= 0.8667



train_loss=0.3256
val_loss=0.6208



TN=15, TP=37, FP=5, FN=3



AUC=0.870

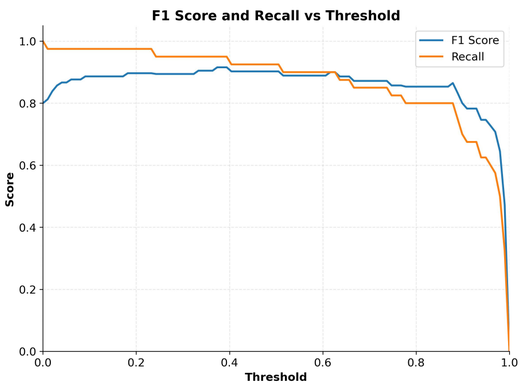


Figure 5: Training Resulets Summary of BACH Dataset

Conclusion

- Developed a hybrid **EfficientNetV2 + ViT** model for breast cancer detection
- Enhanced spatial and contextual learning via **SPT + LSA**
- Achieved high sensitivity in detecting malignant histopathological patterns
- Demonstrated **interpretable predictions** via Grad-CAM heatmaps

Limitation

- Relies solely on histopathological images (**single modality**)
- **Computationally heavy** ViT may hinder deployment in low-resource settings
- Grad-CAM provides basic interpretability, but lacks **uncertainty-aware insights**

Future Work

- Extend to 8-class classification by merging BreakHis & BACH full datasets
- Incorporate clinical metadata and other modalities (e.g., radiology, genomics)
- Explore lighter ViT variants (e.g., MobileViT, Swin) for real-time deployment
- Enhance interpretability with attention trajectory maps or uncertainty visualization

Reference

[1] M. Tan and Q. V. Le, 'EfficientNetV2: Smaller Models and Faster Training', 2021, doi: 10.48550/ARXIV.2104.00298.
[2] S. H. Lee, S. Lee, and B. C. Song, 'Vision Transformer for Small-Size Datasets', 2021, *arXiv*. doi: 10.48550/ARXIV.2112.13492.