Breast Cancer Prediction using Deep Learning

JULIUS CHANG

Compiled March 18, 2025

This project presents a deep learning approach for detecting invasive ductal carcinoma (IDC) in breast histopathology patches. I developed both a custom convolutional neural network (CNN) and a transfer learning model based on ResNet18 to classify 277,524 image patches extracted from whole slide images of over 280 patients. To address class imbalance, on-the-fly data augmentation was applied exclusively to the minority (IDC-positive) class and oversampling was used to equalize training class distributions. The models were evaluated using accuracy, F1-score, ROC AUC, and PR AUC. Our results demonstrate that while the custom CNN exhibited significant overfitting, the transfer learning model achieved robust generalization with validation accuracy exceeding 85% and balanced performance metrics.

1. INTRODUCTION

Breast cancer remains one of the leading causes of morbidity and mortality among women worldwide, with invasive ductal carcinoma (IDC) accounting for approximately 80% of cases [6]. IDC originates in the milk ducts, where malignant cells break through the duct walls and invade surrounding tissue, often leading to metastasis. This aggressive behavior underscores the critical need for early and accurate diagnosis. However, manual histopathological evaluation by pathologists is both time-consuming and subject to variability between observers, potentially delaying treatment decisions. Recent advances in deep learning have shown promising results in medical image analysis, offering the potential to improve diagnostic accuracy and reduce the workload on clinicians. By leveraging these techniques, automated systems can provide consistent, rapid assessments that support and enhance clinical decision-making.

Motivated by these challenges, I explored a deep learning approach to the automated detection of IDC using histopathology patches. My work focused more on the investigation of IDC detection rather than the general breast tissue analysis. Key limitations such as class imbalance—where IDC-positive samples are underrepresented—are addressed by implementing targeted data augmentation and oversampling strategies, applying heavier on-the-fly augmentation exclusively to the minority class. The primary objective is to develop and evaluate a convolutional neural network (CNN) capable of distinguishing IDC-positive from IDC-negative tissue patches. Transfer learning and customized oversampling are used to enhance the model's generalization, thereby improving performance metrics such as precision, F1 score, ROC AUC, and PR AUC. This research aims to provide a robust, expert-independent diagnostic tool that could support clinical decision-making and improve patient outcomes.

2. DATASET

The dataset, sourced from Kaggle, comprises 277,524 image patches extracted from whole slide images of over 280 patients[5][2]. Originally organized by patient and class, the dataset was flattened into two directories corresponding to IDC-positive and IDC-negative patches. Notably, the data exhibits a class imbalance with approximately 28% IDC-positive and 72% IDC-negative patches. To mitigate this, oversampling was implemented and applied heavier augmentation solely to IDC-positive samples during training.

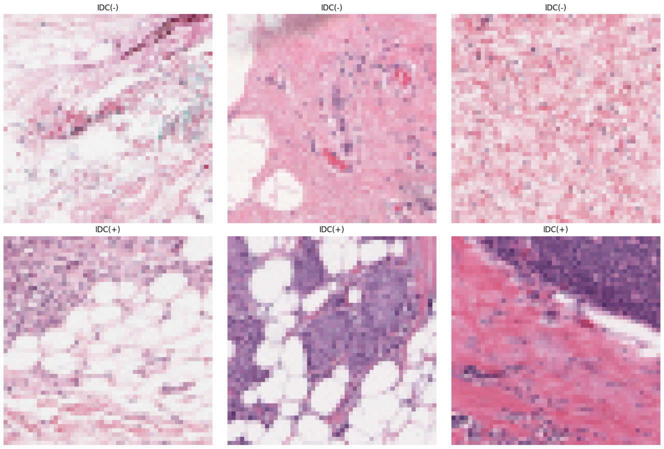

Fig. 1. samples from the dataset of IDC(-) and IDC(+)

3. METHODS

A. Model Architecture

I investigated two approaches: a custom CNN and a transfer learning model using ResNet18. The custom CNN consists of five convolutional layers with batch normalization, dropout, and pooling, followed by five fully connected layers. In contrast, the ResNet18 model leverages pretrained ImageNet weights, with its final fully connected layer replaced to output two classes. For fine-tuning, the earlier layers were frozen and only the final block (layer4) and the new classifier were trained. This strategy was chosen to reduce overfitting while retaining robust feature extraction from the pretrained layers. In addition, a deeper model was utilized, a transformer-based model (Conch ViT-B-16)[1]whose performance will be compared in the Discussion.

B. Training Procedure

Models were trained using the Adam optimizer (learning rate = 0.0001, with weight decay) on a data split of 60% training, 20% validation, and 20% test. To counteract class imbalance in the training set, positive samples were oversampled and augmented using a heavier transformation pipeline, while negatives and validation/test images received basic preprocessing. Training was performed over 30 epochs with learning rate scheduling via a StepLR scheduler. In later experiments, alternative optimizers and learning rate schedulers (e.g., cyclical learning rates, cosine annealing) were considered to further improve convergence. Performance was evaluated using accuracy, F1 score, ROC AUC, and PR AUC.

C. Software and Hardware Environment

Experiments were implemented in Python using PyTorch and torchvision[7]. Data handling was conducted with NumPy and Pandas, and visualizations were generated with Matplotlib. Training was accelerated on an NVIDIA GPU with CUDA support via Google Colab.

4. RESULTS

The experiments indicate that the custom CNN achieved a high training accuracy (up to 98%) but suffered from significant overfitting, with validation and test accuracies around 87–88%(Fig.2). In contrast, the ResNet18 model, fine-tuned with transfer learning, maintained a close match between training and validation performance (both stabilizing around 85–86%)(Fig. 3). Moreover, the ResNet18 model achieved balanced metrics, including an F1 score of approximately 0.72, ROC AUC of 0.90, and PR AUC of 0.81, confirming its superior generalization compared to the custom architecture.The Conch ViT-B-16 model, pretrained on a large histopathology-specific image-caption dataset and integrated with a custom classifier head, exhibited a stable training process over 17 epochs. Final training and validation losses converged around 0.265 and 0.264(Fig. 4), respectively, suggesting minimal overfitting. The validation accuracy stabilized at approximately 88.7–88.8%(Fig. 5), while the test accuracy reached 88.65%, indicating strong generalization to unseen data. Notably, the F1 score (0.79), ROC AUC (0.945), and PR AUC (0.88) are all higher than those observed with our custom CNN and comparable or superior to the pretrained ResNet18 model.

Additionally, the experiments with the transformer-based Conch ViT-B-16 model (details discussed in the supplementary materials) further improved key metrics, indicating that deeper models can capture more complex features, although at the cost of increased computational requirements. These findings underscore that leveraging pretrained models and careful fine-tuning strategies can effectively mitigate overfitting and enhance the robustness of IDC detection, even under constraints such as limited GPU resources on Google Colab.

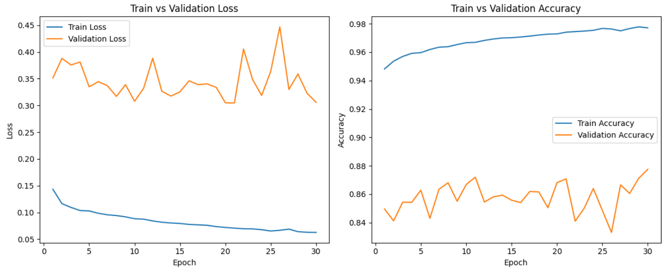

Fig. 2. custom CNN loss and accuracy

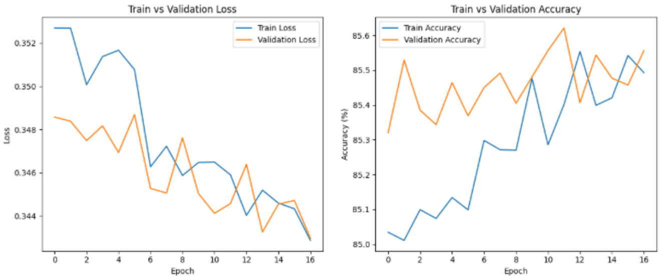

Fig. 3. ResNet18 loss and accuracy

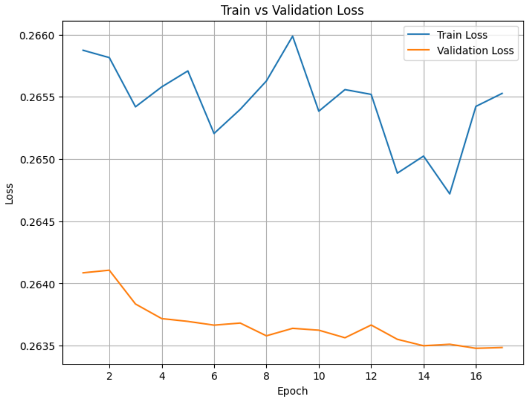

Fig. 4. CONCH Train vs Validation Loss

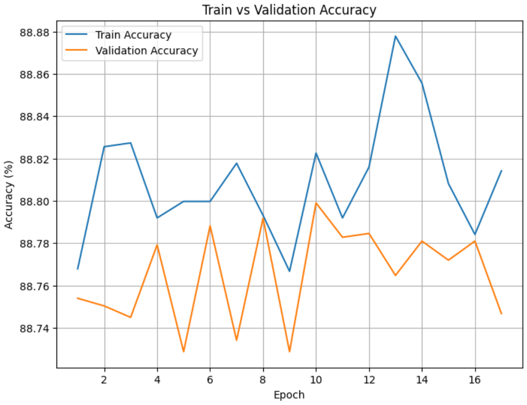

Fig. 5. CONCH Train vs Validation Accuracy

5. DISCUSSION

The study evaluated three models for IDC detection: a custom CNN, a pretrained ResNet18, and a vision-language model based on the CONCH (CONtrastive learning from Captions for Histopathology) architecture. The custom CNN, designed with five convolutional and five fully connected layers, achieved high training accuracy (98%) but suffered from significant overfitting, with validation and test accuracies around 87–88%. In contrast, the ResNet18 model—fine-tuned by unfreezing its final block—yielded more balanced performance, with both training and validation accuracies stabilizing in the mid-80% range and robust metrics (F1 0.72, ROC AUC 0.90, and PR AUC 0.81).

The CONCH-based model leverages a ViT-B-16 vision encoder pretrained on a large histopathology-specific image-caption dataset and is integrated with a custom classifier head[1]. This model benefits from advanced feature representations learned via contrastive learning[1], achieving competitive performance with reduced overfitting. Its metrics—an F1 score around 0.79 and ROC AUC nearing 0.945—suggest that it captures subtle morphological cues of IDC more effectively.

The superior performance of the CONCH model highlights the value of specialized, domain-specific pretrained models in computational pathology. Nevertheless, its increased computational complexity necessitates access to more powerful GPUs than those available on Google Colab. Future work should focus on further mitigating overfitting through enhanced data augmentation, stronger regularization, and the exploration of alternative optimizers and learning rate schedulers. Additionally, expanding the dataset and conducting external validation on independent cohorts will be essential to confirm the generalizability of these promising approaches. Overall, while the custom CNN served as a useful baseline, both the ResNet18 and CONCH models offer more robust and generalizable solutions for IDC detection.

REFERENCES

1. M.-Y. Lu, B. Chen, D. F. K. Williamson, *et al.*, “A visual-language foundation model for computational pathology,” *Nat. Med.*, vol. 30, pp. 863–874, 2024. DOI: 10.1038/s41591-024-02856-4.
2. A. Janowczyk and A. Madabhushi, “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,” *J. Pathol. Inform.*, vol. 7, no. 1, p. 29, 2016. DOI: 10.4103/2153-3539.186902.
3. A. J. Shephard, M. Jahanifar, R. Wang, M. Dawood, S. Graham, and K. Sidlauskas, “An automated pipeline for tumour-infiltrating lymphocyte scoring in breast cancer,” in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, Athens, Greece, May 2024. DOI: 10.1109/ISBI56570.2024.10635302.
4. F. P. Romero, A. Tang, and S. Kadoury, “Multi-level batch normalization in deep networks for invasive ductal carcinoma cell discrimination in histopathology images,” in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, Venice, Italy, Apr. 2019. DOI: 10.1109/ISBI.2019.8759410.
5. Kaggle Contributors, “Invasive Ductal Carcinoma (IDC) Breast Cancer Dataset,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images/data. Accessed: Mar. 17, 2025.
6. Johns Hopkins Medicine, “Invasive Ductal Carcinoma (IDC),” Hopkins Medicine, 2024. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/invasive-ductal-carcinoma-idc. Accessed: Mar. 17, 2025.
7. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *Proc. NeurIPS Workshop (NIPS-W)*, 2017.