**Project Title:**  
Breast Cancer Classification Using Convolutional Neural Networks

**Your Name:**  
Jathin Lalith. D

**Date:**  
7 July 2025

**Abstract**

Breast cancer is a leading cause of mortality among women worldwide. Early and accurate detection is critical to improving patient outcomes. In this project, I developed CancerNet, a Convolutional Neural Network (CNN)-based classifier capable of distinguishing between benign and malignant histology images from the Invasive Ductal Carcinoma (IDC) dataset. The model was trained and evaluated on 277,524 image patches, achieving 85.76% classification accuracy on the test dataset. Results showed the usefulness of applying deep learning for automated histopathological diagnosis.

**Table of Contents**

* Introduction
* Literature Review
* Problem Statement
* Data Collection and Preprocessing
* Methodology
* Implementation
* Results
* Discussion
* Conclusion
* Questions
* References
* Appendices
* Acknowledgments

**1. Introduction**

Breast cancer is among the most common and deadly diseases affecting women worldwide. Early detection is crucial for improving patient survival rates. Traditionally, diagnosing cancer relies on pathologists manually analyzing histopathology slides, which is time-consuming and prone to human error.

This project addresses these challenges by developing CancerNet, a Convolutional Neural Network (CNN) model to automatically classify breast cancer histology images as benign or malignant using the IDC dataset, which contains labeled image patches of breast tissue.

The aim is to create a robust AI system that can help pathologists identify cancer early.

**Objectives:**

* Design and train a CNN model capable of learning key patterns in histology images.
* Evaluate performance using metrics like accuracy, precision, recall, and F1-score.
* Demonstrate the potential of deep learning for improving cancer detection workflows.

**AI Techniques Used:**  
The project uses CNNs implemented with TensorFlow and Keras, along with data augmentation and preprocessing methods to enhance the model’s accuracy.

**2. Literature Review**

Several studies have applied machine learning and deep learning to histopathological image analysis. CNNs have shown state-of-the-art performance in image classification tasks due to their ability to learn hierarchical representations of spatial features.

**Relevant works include:**

* Spanhol et al. (2016), who created the IDC dataset used in this project.  
  <https://www.researchgate.net/publication/283513314_A_Dataset_for_Breast_Cancer_Histopathological_Image_Classification>
* Cireşan et al. (2013), demonstrating CNN efficacy in mitosis detection in breast cancer histology.  
  <https://www.researchgate.net/publication/260432401_Mitosis_Detection_in_Breast_Cancer_Histology_Images_with_Deep_Neural_Networks>
* Recent advances with ResNet and Vision Transformers, which outperform traditional CNNs in some contexts.

**Challenges:**

* Some images were of poor quality.
* Training and testing the model took a long time.

**3. Problem Statement**

**Problem Definition:**  
Develop a Convolutional Neural Network to classify histology images as benign or malignant using the IDC dataset.

**Assumptions & Limitations:**

* The model assumes that 50×50 patches adequately represent cancerous regions.
* Limited to binary classification of IDC presence.
* Performance can be affected by data imbalance.
* Training may take a long time.
* The dataset contains no bad images.

**4. Data Collection and Preprocessing**

**Dataset:**  
IDC\_regular dataset from Kaggle, consisting of 277,524 image patches extracted from breast cancer histology slides.

**Steps Taken:**  
a) Downloaded and verified dataset integrity.  
b) Split data into:

* 80% Training
* 10% Validation
* 10% Testing  
  c) Applied preprocessing:
* Rescaling pixel values
* Data augmentation (rotation, flipping, scaling) for training data
* Normalization
* Organized images into directory structures suitable for Keras generators

**5. Methodology**

**Technique Used:**  
Deep Convolutional Neural Network (CNN)

**Rationale:**  
CNNs are proven efficient and effective for image classification due to their capacity to learn spatial hierarchies and pattern recognition.

**Key Parameters:**

* Input Size: (48×48×3)
* Optimizer: Adam
* Loss: Categorical Crossentropy
* Metrics: Accuracy
* Batch Size: 32
* Epochs: 3 (initial training)

**6. Implementation**

**Model Architecture / Structure:**  
The CancerNet CNN includes:

* Convolutional layers with ReLU activations
* Batch Normalization layers
* MaxPooling layers
* Dropout for regularization
* Flattening and Dense layers with softmax output

**Environment:**

* Frameworks: TensorFlow and Keras
* Language: Python 3.x
* Interface: Jupyter Notebook

**Training and Testing Split:**

* Training: 80%
* Validation: 10%
* Testing: 10%

**Code Repository:**  
<https://github.com/Jathin-24/Project_2>

**7. Results**

**Performance Metrics (after 3 epochs):**

* Accuracy: 85.76%
* Specificity (True Negative Rate): 66.92%
* Sensitivity (True Positive Rate): 93.11%

**Confusion Matrix:**

|  | **Predicted Benign** | **Predicted Malignant** |
| --- | --- | --- |
| Actual Benign | 39,163 | 2,897 |
| Actual Malignant | 5,433 | 10,991 |

**Classification Report:**

| **Class** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 (Benign) | 0.88 | 0.93 | 0.90 | 42,060 |
| 1 (Malignant) | 0.79 | 0.67 | 0.73 | 16,424 |

* Macro Avg F1-score: 0.81
* Weighted Avg F1-score: 0.85

**Training History:**  
Below is a plot of training and validation loss and accuracy across epochs (not included here).

**Comparison to Baselines:**  
Traditional classifiers such as SVMs and Random Forests on similar datasets typically achieve ~75–80% accuracy. CancerNet shows improved performance (~86%) but does not yet reach the accuracy of deeper architectures like ResNet (90%+).

**8. Discussion**

Overall, the model performed well, with high recall for benign cases (about 93%) and good precision (around 88%). However, the recall for malignant cases was lower (67%), indicating the model sometimes misses cancerous samples, which is critical because false negatives can be serious.

There is not much overfitting, likely due to training for only 3 epochs. Longer training (5–10 epochs) might improve results. Using more advanced models or transfer learning with pretrained networks like ResNet could increase accuracy, especially for malignant cases.

Strength: CNNs effectively learn image features automatically.  
Limitation: Sensitivity to cancer cases was lower than desired, indicating room for improvement in detecting malignant patches.

**9. Conclusion**

This project successfully implemented a CNN classifier, CancerNet, capable of distinguishing benign and malignant breast cancer histology patches with ~86% accuracy. This demonstrates the potential of deep learning to assist pathologists in early diagnosis and treatment planning.

**Potential Real-World Use:**

* Pre-screening histology slides to prioritize cases for pathologist review.
* Supporting early detection efforts in hospitals and research centers.
* Integrating into automated pathology workflows to improve consistency and reduce workload.

**Future Work:**

* Training additional epochs (5, 10, or more) to improve convergence.
* Incorporating explainability methods.
* Experimenting with transfer learning for higher accuracy.

**10. Questions**

**What is the training and testing split you used?**  
80% training, 10% validation, 10% testing.

**How many epochs did you run your model?**  
3 epochs initially.

**Is CNN the best model? Alternatives?**  
CNNs are strong for images, but ResNet or Vision Transformers may perform better.

**Accuracy after 5 or 10 epochs?**  
Not measured yet; likely higher than the 3-epoch result of ~86%.

**Is your model overfitting or underfitting?**  
No clear overfitting observed in 3 epochs; further training recommended.

**How would you use it in real life?**  
Integrate into pathology labs as a triaging tool to assist pathologists.

**11. References**

* Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455–1462.
* Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction  
  <https://pmc.ncbi.nlm.nih.gov/articles/PMC9818155/>
* Kaggle Breast Cancer Detection Dataset  
  <https://www.kaggle.com/code/prakharbhartiya1/breast-cancer-detection/data>

**12. Appendices**

**A. Model Code Snippets**  
Full implementation is provided in the accompanying Jupyter Notebook.

**B. Instructions to Reproduce**

* Install dependencies:  
  pip install tensorflow keras imutils matplotlib
* Download the IDC dataset.
* Open the notebook and execute all cells sequentially.

**13. Acknowledgments**

Special thanks to the TensorFlow community and the Kaggle platform for providing valuable resources and datasets that made this project possible.