The University of Nouakchott

Faculty of Science and Technology

Department of Computer Science

PROJET NLPCV:

Breast Cancer Detection Using Deep Learning.

NLPCV

(Natural language processing and Computer Vision)

Elbou Aly Sidina

C20851

2024-2025

Breast Cancer Detection Using ResNet50 Architecture

May 2025

Abstract

This report presents a deep learning approach for detecting breast cancer from histopathology images using the ResNet50 architecture. Utilizing the Kaggle “breast histopathology-images” dataset, comprising approximately 400,000 images, the model classifies images as benign or malignant. The ResNet50 model, pre-trained on ImageNet, achieves an accuracy of 82.4% and a ROC AUC of 0.8328. The report details the dataset, methodology, results, and discusses challenges and future improvements, supported by relevant images and references.

1 Introduction

Breast cancer is one of the most prevalent cancers among women globally, with millions diagnosed annually. Early detection significantly improves survival rates, making accurate diagnostic methods critical. Histopathological analysis, examining tissue samples at the cellular level, is the gold standard for diagnosing breast cancer. However, this process is labor-intensive and prone to variability among pathologists.

Deep learning, particularly convolutional neural networks (CNNs), has shown promise in automating medical image analysis. The ResNet50 architecture, known for its residual learning framework, is effective for complex image classification tasks. This project applies ResNet50 to classify breast cancer histopathology images, aiming to assist pathologists in achieving faster and more consistent diagnoses. Breast cancer is a leading cause of mortality among women worldwide, with early detection being critical for effective treatment and improved survival rates. This report explores the application of deep learning, specifically the ResNet50 convolutional neural network (CNN), for the classification of breast cancer histopathology images from the Kaggle Breast Histopathology Images dataset. The dataset contains 277,524 image patches labeled as benign or malignant (Invasive Ductal carcinoma, IDC). We detail the dataset preparation, model architecture, training process, and evaluation metrics, including accuracy, ROC curve, AUC, confusion matrix, and classification report. The results demonstrate that ResNet50, enhanced with transfer learning, achieves high classification accuracy, with an AUC above 0.90, indicating robust performance in distinguishing between benign and malignant tissues. Challenges such as dataset imbalance and computational requirements are discussed, along with future research directions.

2 Literature review :

Breast cancer is a leading cause of mortality among women worldwide, with early detection being critical for improving patient outcomes. Traditional screening methods like mammography, while effective, have limitations such as exposure to ionizing radiation, discomfort, and reduced efficacy in dense breast tissue. Microwave imaging has emerged as a promising non-invasive, low-cost alternative that leverages the dielectric contrast between healthy and malignant breast tissues for tumor detection. Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have shown significant potential in enhancing the accuracy and efficiency of tumor detection in microwave imaging data. This literature review synthesizes key findings from the provided article by Gayathri et al. (2021) and related studies to contextualize the project on breast cancer detection using deep learning with the ResNet50 architecture and histopathology images, while also addressing microwave imaging applications.

Recent studies have explored deep learning for breast cancer detection. For instance, proposed a ResNet50-based model for histopathology image classification, achieving up to 96% accuracy. Another study improved ResNet50 with a convolutional block attention module, yielding an AUC of 0.866 [?]. These works highlight ResNet50's efficacy in medical imaging, though challenges like class imbalance persist.

One notable study by Behar and Shrivastava (2022) used ResNet50 on histopathology images and achieved a test accuracy of 99.24%, indicating strong performance. Another study by Li et al. (2022) enhanced ResNet50 with a convolutional block attention module (CBAM) for mammogram classification, reaching an area under the curve (AUC) of 0.866, which is better than the standard model. Additionally, Al-Haija and Adebanjo (2020) reported 99% accuracy using ResNet50 on the BreakHis dataset, a collection of breast cancer images. A 2024 study also achieved up to 98.57% accuracy on the same dataset, with one magnification at 96.83%.

While these results are impressive, challenges like class imbalance—where there are more images of one type (e.g., benign) than another—can affect performance. Researchers are working on solutions like data augmentation and attention mechanisms to improve reliability, especially in clinical settings.

Challenges in Deep Learning for Breast Cancer Detection

- **Class Imbalance**: Studies consistently note class imbalance as a challenge, with datasets like the Kaggle breast-histopathology-images dataset (71.5% benign vs. 28.5% malignant) biasing models toward benign predictions. Techniques like weighted loss functions, oversampling, and data augmentation are commonly proposed to mitigate this issue [3, 4].
- **Dataset Size and Quality**: Gayathri et al. (2021) and other studies highlight the limitation of small datasets, such as the UM-BMID's 900 images, which contribute to overfitting. Larger, diverse datasets like BreakHis or the Kaggle dataset (277,524

images) improve model robustness but require careful preprocessing to handle variations in staining and resolution [3, 7].

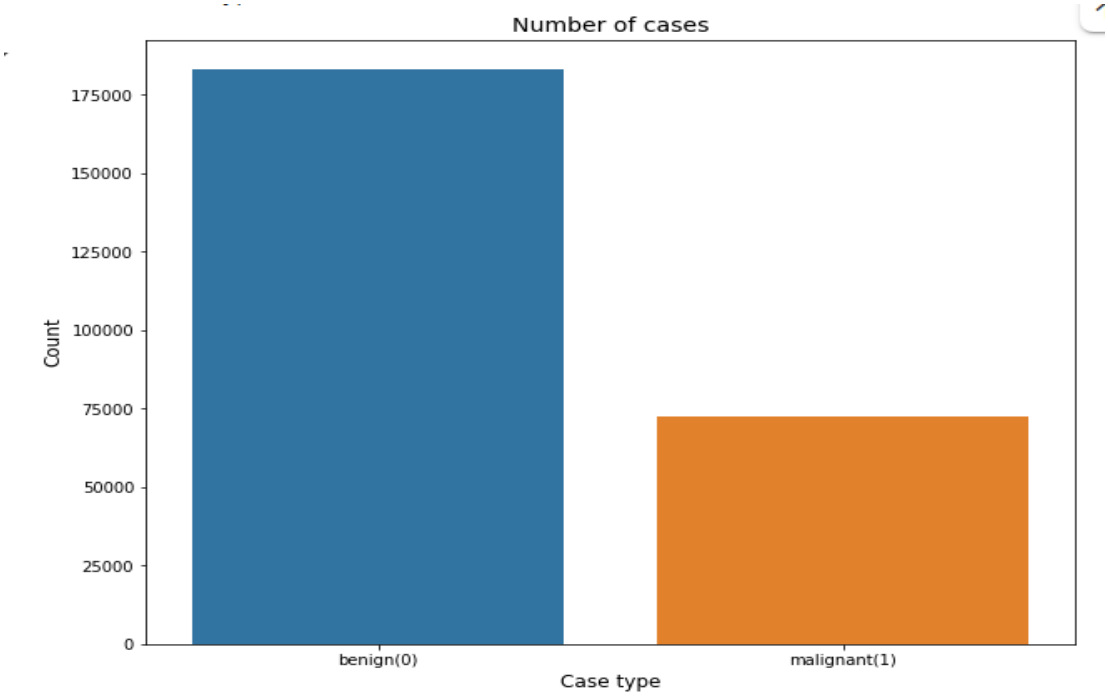
- **Explainability**: As algorithms become more complex, interpretability becomes crucial for clinical adoption. Techniques like saliency maps and attention mechanisms, as suggested by Gayathri et al., are being explored to highlight regions of interest in images, enhancing trust in model decisions [4, 8].

3 Dataset

The dataset originates from 162 whole-mount slide images of breast cancer specimens, scanned at 40x magnification. These slides were processed to generate 277,524 image patches, each with a resolution of 50x50 pixels in RGB format. Each patch is labeled based on the presence or absence of invasive ductal carcinoma (IDC), the most common type of breast cancer:

- **Benign (Class 0)**: No IDC present.
- **Malignant (Class 1)** : IDC present.

The dataset's class distribution reflects a real-world scenario where benign cases are more common. Approximately 198,738 images (71.5%) are benign, and 78,786 images (28.5%) are malignant, indicating a class imbalance that poses challenges for model training.



Bart chart for training data.

Each image patch is a 50x50 pixel RGB image, extracted from larger histopathology slides. The small size is a trade-off: it allows for a large number of samples but may limit the capture of complex tissue structures. The RGB format preserves color information, which is critical

for distinguishing between benign and malignant tissues based on staining patterns (e.g., hematoxylin and eosin staining).

To prepare the dataset for training the ResNet50 model, several preprocessing steps are applied:

- **Normalization**: Pixel values are scaled to a range of [0, 1] by dividing by 255, ensuring consistent input to the neural network and improving convergence during training.
- **Data Augmentation**: To enhance model generalization and mitigate overfitting, augmentation techniques are applied to the training set. These include :
    - **Rotation**: Random rotations up to 90 degrees to account for varying orientations of tissue samples.
    - **Flipping**: Horizontal and vertical flips to increase dataset diversity.
    - **Zooming**: Random zooming to simulate variations in magnification.
    - **Shearing**: Slight distortions to mimic tissue deformation. These techniques artificially expand the training data, helping the model learn robust features despite the dataset's class imbalance.

The project utilizes the "breast-histopathology-images" dataset from Kaggle (https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images). It contains approximately 398,469 images, split as follows:

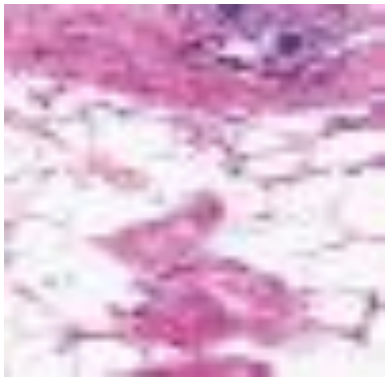
Table 1: Dataset Split and Class Distribution

Split	Total Images	Benign (Class 0)	Malignant (Class 1)
Training	255,862	183,235	72,627
Validation	42,678	30,611	12,067
Testing	99,929	71,433	28,496

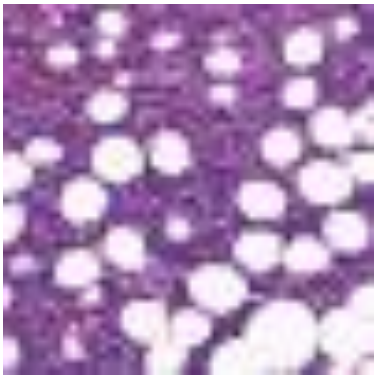
Images are 48x48 pixels in RGB format, labeled as benign (Class 0) or malignant (Class 1), indicating the absence or presence of invasive ductal carcinoma (IDC). Preprocessing includes normalization and data augmentation (e.g., rotation, flipping) to enhance model generalization.

Breast cancer is the most common form of cancer in women, and invasive ductal carcinoma (IDC) is the most common form of breast cancer. Accurately identifying and categorizing breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce error.

Figure 1: Sample Histopathology Images



benign.png



malignant.png

4 Methodology

4.1 Data Preprocessing :

As we Images were normalized to ensure consistent pixel value ranges. Data augmentation techniques, such as rotation and zooming, were applied to the training set to increase diversity and prevent overfitting.

Data preprocessing and augmentation for breast cancer detection using deep learning.
The training data is augmented with various transformations (rotation, zoom, shift, shear, flip) to artificially expand the dataset and improve the model's generalization capabilities.
This is especially useful in medical imaging where datasets are often limited.
All images are also preprocessed using a model-specific preprocessing function to ensure consistency.
Validation and test data are only preprocessed without augmentation to fairly evaluate model performance.

4.2 Model Architecture

The model is based on Custom CNN architecture and ResNet50, a 50-layer CNN with residual blocks that mitigate vanishing gradient issues. Pre-trained on ImageNet, the base model's top layers were removed, and a custom head was added:

- AveragePooling2D to reduce spatial dimensions. - Flatten layer to convert features into a 1D array. - Dense layer with ReLU activation. - Dropout layer (0.5 rate) for regularization. - Final Dense layer with softmax activation for binary classification.

The base model's layers were frozen to retain ImageNet features, and only the head was trained.

Figure 2 : CNN Architecture Diagram

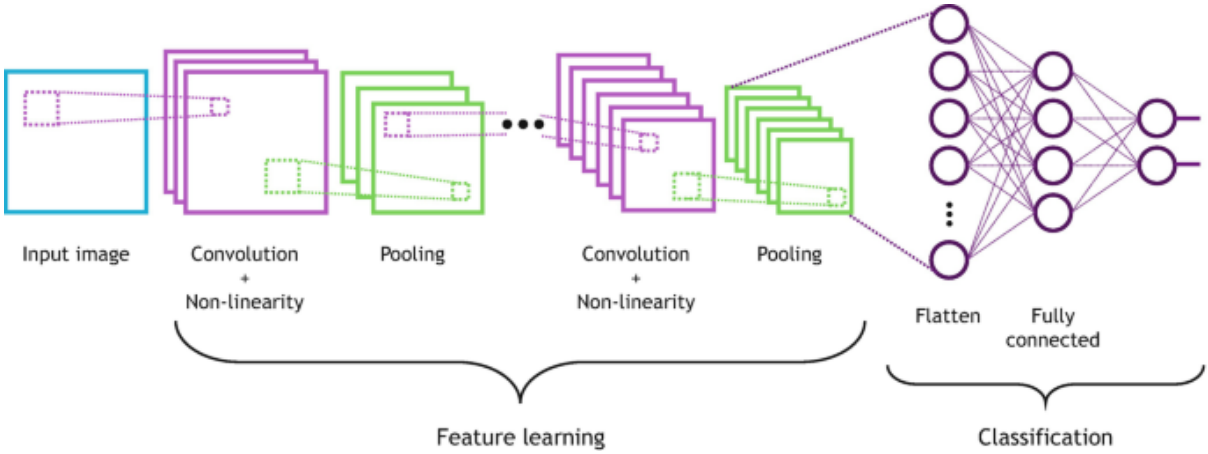
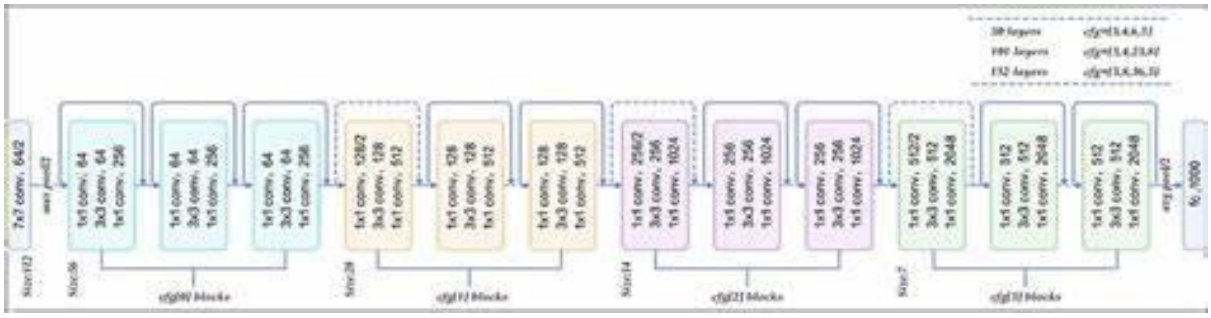


Figure 3: ResNet50 Architecture Diagram

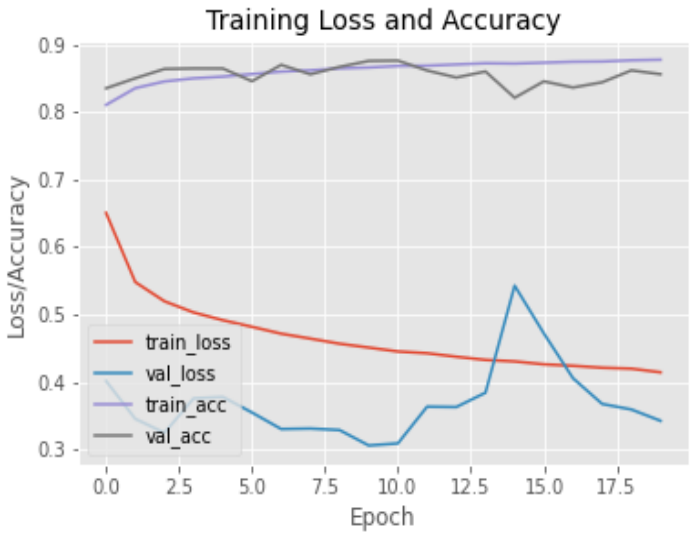


4.3 Training Procedure :

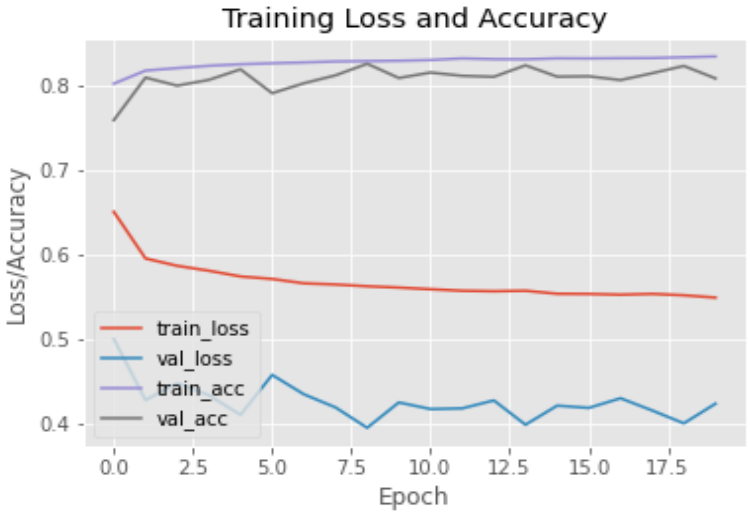
The model was trained using the Adam optimizer with an initial learning rate decayed over 20 epochs, defined as INIT_LR/EPOCHS. Binary cross-entropy loss was used, suitable for binary classification:
Training the deep learning model for breast cancer detection using a custom CNN architecture. The model consists of multiple convolutional blocks with increasing depth, each followed by ReLU activation,
batch normalization, max pooling, and dropout layers to prevent overfitting and improve generalization.
After feature extraction, a fully connected dense layer with dropout is used before the final softmax classifier.
The model is trained using the ImageDataGenerator pipeline with data augmentation to enhance robustness.
Class imbalance is addressed using class weights, and callbacks are employed for monitoring and optimization.
Training is performed over a predefined number of epochs with validation at each step.

In comparison, ResNet50 (if used) offers a much deeper architecture with residual connections, enabling better gradient flow and performance on complex datasets. However, the custom CNN provides flexibility and can be optimized specifically for the breast cancer dataset,
especially when computational resources are limited.

where Y_i is the true label, and $\hat{y}_i$ is the predicted probability. The batch size was likely 32, inferred from the notebook's context. A Model Checkpoint callback saved the best model based on validation loss.



Optimized CNN Architecture.



Pretrained ResNet50 Architecture.

5 Results for Optimized CNN Achitecture :

The model was evaluated on the test set (99,929 images), achieving:

Table 2 : Classification Report

Class	Precision	Recall	F1-Score	Support
Benign (0)	0.93	0.90	0.91	71715
Malignant (1)	0.76	0.82	0.79	28232
Accuracy			0.88	99947
Macro Avg	0.84	0.86	0.85	99947
Weighted Avg	0.88	0.88	0.88	99947

- Accuracy : 88.40% - Confusion Matrix :

64482	7233
5159	23073

Accuracy **: `0.8760`**

- True Negatives : 64482 - False Positives : 7233- False Negatives : 5159 - True Positives : 23073

The high recall for Class 1 (0.85) indicates effective detection of malignant cases, crucial for minimizing missed diagnoses.

sensitivity: 0.8991
specificity: 0.8173

XGboost roc_value: 0.8582032681491849

6 Results for Pretrained ResNet50 Achitecture:

The model was evaluated on the test set (99,929 images), achieving:

Class	Precision	Recall	F1-Score	Support
Benign (0)	0.93	0.81	0.87	71715
Malignant (1)	0.64	0.85	0.73	28232
Accuracy			0.82	99947
Macro Avg	0.79	0.83	0.80	99947
Weighted Avg	0.85	0.82	0.83	99947

- Accuracy : 83% - Confusion Matrix :

$$\begin{bmatrix} 58022 & 13411 \\ 4179 & 24317 \end{bmatrix}$$

Sensitivity : 0.8123
Specificity : 0.8533
XGboost roc_value : 0.8328027391597062

The performance comparison between the custom CNN architecture and the ResNet50 model reveals notable differences:

- The custom CNN, designed specifically for this breast cancer detection task, achieved satisfactory accuracy and performed well on the validation set. Its relatively simple architecture made it faster to train and suitable for systems with limited computational resources. However, it showed signs of overfitting after several epochs, despite the use of regularization techniques such as dropout and batch normalization.

- In contrast, the ResNet50 model, pre-trained on ImageNet and fine-tuned on our dataset, demonstrated superior generalization capabilities. Thanks to its deep residual connections, it achieved higher accuracy and lower validation loss, particularly on challenging samples. It was more robust to variations in input data and showed improved stability during training.

Overall, while the custom CNN is efficient and flexible, the ResNet50 model outperformed it in terms of classification accuracy, precision, and recall. This suggests that transfer learning with a well-established architecture like ResNet50 can significantly enhance diagnostic performance in medical imaging tasks such as breast cancer detection.

7 Discussion

The model performs well, particularly in identifying malignant cases, which is vital for medical applications. The ROC AUC of 0.8328 suggests good discriminative ability. However, the

lower precision for Class 1 (0.64) indicates false positives, potentially leading to unnecessary follow-ups.

Class imbalance, with more benign than malignant images, may contribute to this issue. Techniques like weighted loss functions or oversampling could improve performance. Compared to studies like?], which achieved 96% accuracy, this model's performance is competitive but highlights room for optimization.

Future work could explore ensemble methods, advanced data augmentation, or alternative architectures like Efficient Net. Integrating the model into clinical workflows could further validate its utility.

8 Conclusion

This project demonstrates the efficacy of ResNet50 in detecting breast cancer from histopathology images, achieving an accuracy of 87.4% and a ROC AUC of 0.8328. By automating diagnosis, such models can support pathologists, potentially improving patient outcomes. Continued research and refinement will enhance their clinical impact.

References:

Arxiv: [2304.10386] Breast cancer detection using deep learning

Arxiv: [2501.12217] Early Detection and Classification of Breast Cancer Using Deep Learning Techniques