**Title Page:**

**Project Title:**

**Breast Cancer ------------------Classification**

**----------------------------Yella Harini  
----------------------------------------------------22-08-2024**

**Project Title:**

Breast Cancer Histology Image Classification Using CNN (CancerNet)

**Your Name:**

Yella Harini

**Date:**  
21/8/24

**Abstract:**

Breast cancer remains a significant global health concern, necessitating early detection for improved treatment outcomes. This project focuses on developing an AI-based classifier, **Cancer Net**, using Convolutional Neural Networks (CNN) to classify breast cancer histology images as benign or malignant. Leveraging the IDC dataset, the model is trained to analyze microscopic images of breast tissue and provide accurate diagnoses. With extensive training and testing, the model achieves high accuracy, offering a potential tool for early cancer detection in real-world medical applications.

**Table of Contents:**

1. Introduction

2. Literature Review

3. Problem Statement

4. Data Collection and Preprocessing

5. Methodology

6. Implementation

7. Results

8. Discussion

9. Conclusion

10. References

11. Appendices

12. Acknowledgments (optional)

**Introduction:**

Breast cancer is one of the most common cancers affecting women globally, with early detection being crucial for improving survival rates. This project aims to leverage AI techniques, particularly Convolutional Neural Networks (CNN), to classify breast cancer histology images. The primary objective is to build \*\*CancerNet\*\*, a CNN model capable of distinguishing between benign and malignant breast cancer images with high accuracy. This model can significantly assist pathologists in diagnosing breast cancer more efficiently and could be integrated into healthcare workflows for early detection.

**Literature Review:**

Previous studies have applied machine learning and deep learning techniques to medical imaging for cancer classification. CNNs have been particularly successful due to their ability to extract features from image data. While traditional machine learning models like Support Vector Machines (SVM) and Random Forests have been used, CNNs outperform these methods in image classification tasks. However, challenges remain in obtaining high accuracy with large datasets and avoiding overfitting. The advancements in medical AI have demonstrated the potential for real-world implementation, but data preprocessing and model interpretability remain areas of active research.

**Problem Statement:**

Breast cancer is a leading cause of death worldwide, and early diagnosis is critical to improving patient outcomes. The challenge is to develop a model that can accurately classify breast cancer histology images as benign or malignant. This project will focus on analyzing histopathology images using CNNs. Key challenges include ensuring high accuracy, avoiding overfitting, and optimizing model performance on a relatively imbalanced dataset.

**Assumptions and Limitations:**

- The dataset used contains small image patches (50x50), which may limit feature extraction capabilities.

- Real-world data might contain more noise or variability than what is present in the dataset.

**Data Collection and Preprocessing:**

The dataset used in this project is the \*\*IDC Regular dataset\*\*, containing 277,524 image patches of size 50x50. These patches are extracted from whole mount slide images of breast cancer specimens. Of these, 1,98,738 are benign (test negative), and 78,786 are malignant (test positive).

**Preprocessing Steps:**

- Image normalization (scaling pixel values between 0 and 1).

- Data augmentation (random rotations, flips) to prevent overfitting.

- Splitting the data into training and testing sets (80/20 split).

**Methodology:**

The project employs a Convolutional Neural Network (CNN) due to its effectiveness in handling image data. The CNN architecture, **CancerNet**, consists of several convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification.

**CNN Layers:**

- Convolutional Layers: Extract key features (edges, textures) from the images.

- Pooling Layers: Reduce dimensionality while retaining important features.

- Dropout Layers: Prevent overfitting by randomly dropping connections during training.

- Fully Connected Layers: Perform the final classification (benign or malignant).

**Activation Functions:**

- ReLU for hidden layers to introduce non-linearity.

- Sigmoid for the output layer, as it is a binary classification problem.

**Optimizer**:

Adam optimizer was used for fast convergence with a learning rate of 0.001.

**Implementation**:

The model was implemented using Python and the Keras library. Training was done on the IDC dataset, with the CNN model learning to classify images as benign or malignant.

- **Epochs**: The model was trained over 20 epochs, with results evaluated after every 5 epochs.

- **Loss Function:** Binary Cross-Entropy was used, as the task is binary classification.

You can view the implementation on GitHub: [https://github.com/HariniReddy07/AI-Prj2-IntrnForte]

**Results:**

After training for 20 epochs, the following results were obtained:

- Accuracy after 5 epochs: 80.2%

- Accuracy after 10 epochs: 88.5%

- Final Accuracy: 92.1%

The confusion matrix revealed that the model correctly identified most malignant and benign cases, although there were a few false positives and false negatives. The performance was consistent across both the training and validation datasets, indicating that the model was neither underfitting nor overfitting.

**Discussion:**

The model demonstrates strong performance in classifying breast cancer histology images, with over 90% accuracy. One of the strengths of this model is its ability to generalize well on unseen data due to data augmentation and dropout layers. However, further improvements could be made by experimenting with deeper architectures or using more advanced techniques like Vision Transformers (ViTs).

Unexpectedly, a few benign images were misclassified as malignant, potentially due to overlapping features between certain benign and malignant regions. This indicates a need for further feature engineering or possibly combining CNNs with other models for improved precision.

**Conclusion:**

This project successfully developed a CNN model, **CancerNet**, capable of classifying breast cancer histology images with high accuracy. The model has the potential to assist medical professionals in early cancer detection, reducing the time for diagnosis and improving patient outcomes. Future work could involve testing the model on larger datasets or incorporating it into medical diagnostic systems for real-time predictions.

**References:**

1. [Kaggle - IDC Dataset](https://www.kaggle.com/)

2. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. \*Nature\*, 521(7553), 436-444.

3. Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. \*Medical image analysis\*, 42, 60-88.

**Appendices:**

- Appendix A: Confusion Matrix

- Appendix B: Code Snippets for Model Training and Evaluation (https://github.com/HariniReddy07/AI-Prj2-IntrnForte)

**Acknowledgments:**

I would like to thank [Individuals, Institutions, Mentors] for their guidance and support throughout this project. Special thanks to [Data Providers] for making the IDC dataset publicly available.