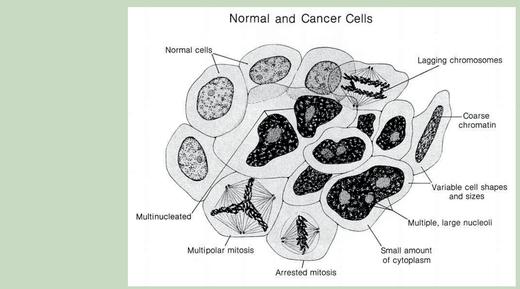
Prediction Of Breast Cancer Incidence

Shivani - 23200782

Scan me!

01 Introduction

Breast cancer is a leading cause of mortality among women globally, making early detection crucial for effective treatment. This project leverages machine learning and deep learning techniques to enhance the accuracy of breast cancer detection. By utilizing the Wisconsin Breast Cancer dataset and the BreaKHis 400X dataset, the project focuses on classifying breast tumor images and analyzing cell nuclei characteristics to predict malignancy.



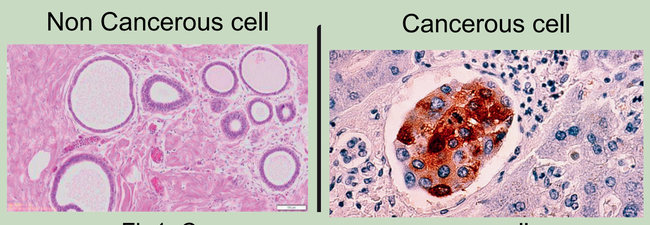


Fig1: Cancerous vs non-cancerous cell

02 Objective

To develop and evaluate machine learning models that accurately classify breast cancer images and predict malignancy using the Wisconsin Breast Cancer and BreaKHis 400X datasets.

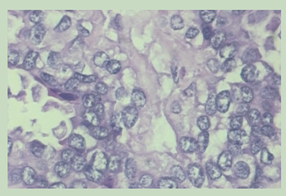
To enhance early detection capabilities by analyzing cell nuclei characteristics and utilizing deep learning techniques for improved diagnosis and treatment planning.

03 Exploratory Data Analysis

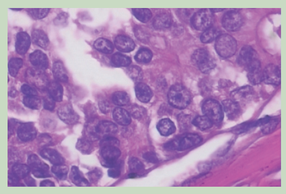
To aid in the fulfilling the set goals, there are two datasets used in the detection, prediction and classification of breast cancer incidence, and they are:

1. BreaKHis 400X Dataset:
   a. The BreaKHis 400X dataset contains high-resolution microscopic images of breast tumour tissues, categorized into benign and malignant classes at 400x magnification.
   b. It is used to train deep learning models for distinguishing between benign and malignant breast cancer based on visual characteristics of cell structures.

1. Breast Cancer Data - Wisconsin:
   a. The Breast Cancer Data - Wisconsin dataset consists of numeric features derived from fine needle aspirates of breast masses, used for predicting breast cancer diagnosis.
   b. It includes measurements like radius, texture, and perimeter of cell nuclei, serving as input for machine learning models to classify tumors as benign or malignant.



Benign cell under a microscope



Malignant cell under a microscope

Fig2: Cells under a microscope

Fig3: EDA of Target varaible and features

04 Methodology

Data Splitting:
Both datasets were split into 80% for training and 20% for testing to ensure robust model evaluation.

Model Selection:
For the BreaKHis dataset, a DenseNet CNN architecture was employed to train and classify the images. The Wisconsin dataset was analyzed using Random Forest, Logistic Regression, KNN, and Neural Networks.

Model Evaluation:
The DenseNet CNN achieved 96.4% training accuracy and 87.45% test accuracy. Random Forest outperformed other models on the Wisconsin dataset with superior accuracy, precision, recall, and F1 score.

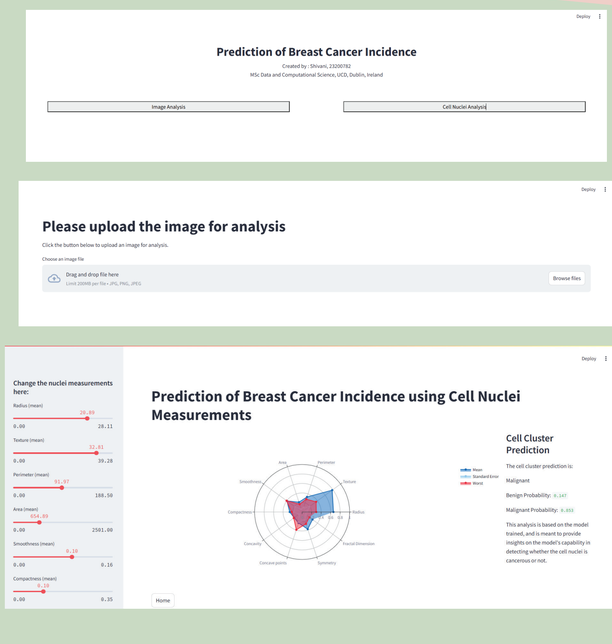


Fig4: Illustration of the UI

05 Results

The DenseNet model's performance on the BreaKHis dataset demonstrated strong classification capabilities. The training accuracy reached 96%, with validation accuracy showing stability across epochs. The evaluation metrics are impressive, with an accuracy of 96%, precision of 0.91, recall of 0.94, and an F1 score of 0.97. The confusion matrix further validates the model's effectiveness, correctly classifying 150 benign and 358 malignant cases, with minimal misclassifications (26 benign as malignant and 11 malignant as benign). The loss and accuracy graphs indicate consistent training with minimal overfitting, confirming the model's reliability(Fig7).

The performance of various models on the Wisconsin Breast Cancer dataset was evaluated and compared (Fig6). The Random Forest model outperformed the others with an accuracy of 96.49%, F1 score of 0.952, and a balanced accuracy of 0.958. The Neural Network and Logistic Regression models also demonstrated high performance with accuracies of 95.61%. K-Nearest Neighbors, while accurate, showed slightly lower performance with an accuracy of 94.73%. These metrics highlight the effectiveness of Random Forest in accurately classifying breast cancer instances based on the dataset.

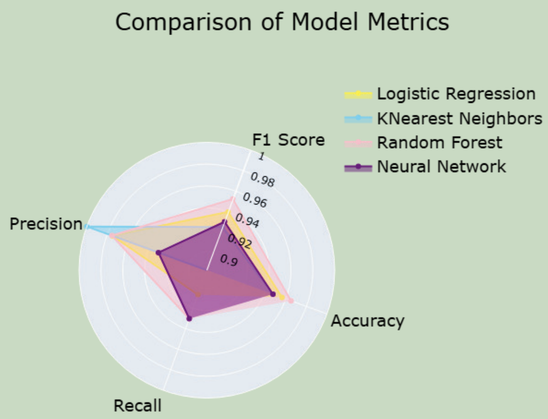
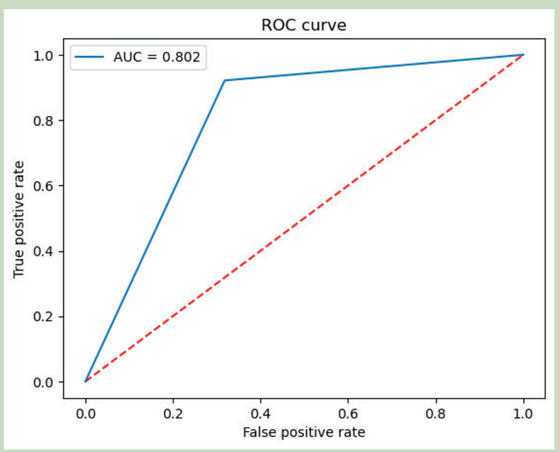


Fig6: Evaluation metrics of various models



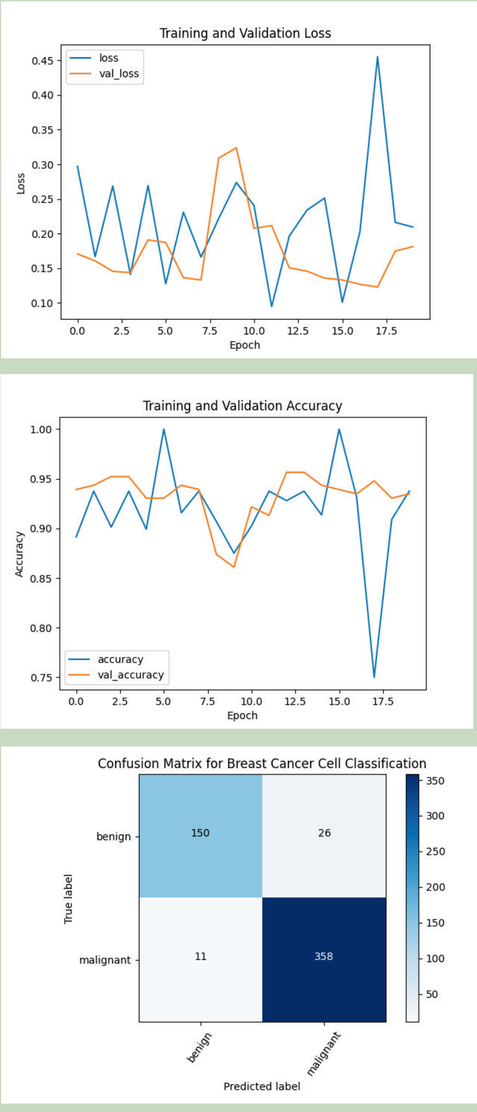


Fig7: Results of DenseNet CNN

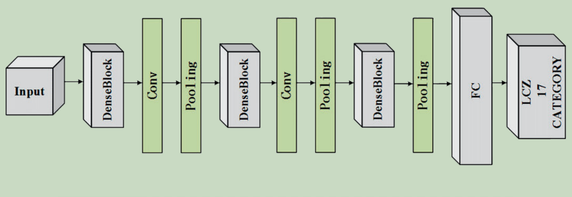


Fig5: DenseNet Architecture

06 Conclusion & Future Work

This project demonstrates the potential of machine learning and deep learning models in accurately predicting breast cancer. The DenseNet-based CNN achieved a commendable training accuracy of 96.4% on the BreaKHis dataset, and the Random Forest model emerged as the best-performing model on the Wisconsin dataset with an accuracy of 96.49%. These results highlight the efficacy of advanced algorithms in distinguishing between benign and malignant cases, providing valuable support in medical diagnosis.

However, the variability in model performance across different datasets emphasizes the importance of dataset-specific model selection and further model refinement for clinical applications.

Future work could focus on integrating more diverse datasets and exploring ensemble methods to enhance model robustness and generalization.

07 References

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S. et al. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040, The Breast 66: 15–23.

Huang, J., Chan, P. S., Lok, V., Chen, X., Ding, H., Jin, Y., Yuan, J., Lao, X.-q., Zheng, Z.-J. and Wong, M. C. (2021). Global incidence and mortality of breast cancer: a trend analysis, Aging (Albany NY) 13(4): 5748.

Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M. A., Gulliver, T. A. and Tamez-Pe˜na, J. G. (2024). Breast cancer risk prediction using machine learning: a systematic review, Frontiers in Oncology 14: 1343627.

Muller, F. M., Li, E. J., Daube-Witherspoon, M. E., Vanhove, C., Vandenberghe, S., Pantel, A. R. and Karp, J. S. (2024). Deep learning denoising for low-dose dual-tracer protocol with 18f-fgln and 18f-fdg in breast cancer imaging, Annual Meeting of the Society of Nuclear Medicine and Molecular Imaging.

FA Spanhol, LS Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016, doi: 10.1109 / TBME.2015.2496264.

Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis
Mohammed Amine Naji , Sanaa El Filali , Kawtar Aarika , EL Habib Benlahmar , Rachida Ait Abdelouhahid , Olivier Debauche.