IDENTIFICATION OF BREAST CANCER FROM HISTOPATHOLOGICAL IMAGES IN KENYA

INTRODUCTION

Breast cancer stands as a formidable public health challenge in Kenya, characterized by alarmingly high mortality rates largely attributed to late-stage detection. The country's healthcare system grapples with limited resources, including a shortage of skilled pathologists and a dearth of advanced diagnostic technologies. Consequently, the prevailing manual diagnostic methods are time-consuming, error-prone, and often lead to delayed or inaccurate diagnoses. This project aims to address these critical issues by developing and implementing a robust machine learning model capable of accurately classifying breast cancer types from histopathological images.

BUSINESS UNDERSTANDING

The overarching goal of this project is to address the critical issue of breast cancer in Kenya, characterized by high mortality rates due to late-stage detection and limited access to quality healthcare. By developing a deep machine learning-based solution, this aims to enhance early diagnosis, improve diagnostic efficiency, and expand access to

accurate breast cancer screening services. Ultimately, this project seeks to contribute to a significant reduction in breast cancer mortality rates and improve the overall quality of life for Kenyan women.

PROBLEM STATEMENT

Breast cancer poses a significant public health challenge, characterized by high mortality rates primarily due to late-stage detection. The country's healthcare system faces numerous obstacles in addressing this crisis, including limited access to specialized healthcare, a shortage of skilled pathologists, and a lack of advanced diagnostic technologies. Consequently, the current manual diagnostic process is time-consuming, error-prone, and often leads to delayed or inaccurate diagnoses. This delay in identifying and treating breast cancer has severe implications for patient outcomes, including increased morbidity and mortality rates. To mitigate these challenges and improve breast cancer care in Kenya, there is a critical need for innovative solutions that can enhance early detection, improve diagnostic accuracy, and increase accessibility to quality care.

OBJECTIVES

1. Develop a robust image classification model: Create an accurate and efficient machine learning model capable of distinguishing between benign and malignant breast tissue based on histopathological images.
2. Improve diagnostic efficiency: Accelerate the diagnostic process by automating image analysis tasks, reducing the workload on pathologists, and enabling faster patient treatment.
3. Enhance patient outcomes: Contribute to early detection of breast cancer, leading to improved treatment options and increased survival rates.

Goals

- Analyze existing data on breast cancer in Kenya.
- Develop a machine learning model for classifying histopathological images.
- Evaluate the model's performance and potential impact.
- Explore integration strategies for the model into Kenya's healthcare system.

METRICS OF SUCCESS

1. Accuracy: As a measure of the proportion of correctly predicted instances out of the total instances. It reflects the model's ability to classify data points accurately, with higher accuracy indicating better performance in distinguishing between different classes.
2. Loss: As the measure of how well the model's predictions match the actual values. It quantifies the difference between predicted and true values, with a lower loss indicating a more accurate model. Evaluating the test loss helps determine how well the model can generalize to new, unseen data.

Expected Outcomes

- A better understanding of the breast cancer landscape in Kenya.
- A machine learning model capable of accurately classifying breast cancer types.
- Recommendations for implementing the model in Kenyan healthcare settings.

Methodology

Data Collection

The BreakHis dataset, a publicly available dataset of histopathological images of breast tissue, was used for this project. The dataset consists of 7,909 images, divided into two classes: benign (2,480 images) and malignant (5,429 images). this is inclusive of the X40, X100, X200, X400 Magnification levels for the main categories and their subcategories.

(https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
(https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/))

Data Preparation & Cleaning

- Eliminate duplicates and irrelevant images to reduce noise and improve model performance.
- Enhance images by filtering, sharpening, or adjusting contrast to improve clarity.
- Crop images to focus on relevant subjects and resize them to a consistent dimension.
- Conduct manual checks to verify the accuracy of images and labels, ensuring quality control.
- Create a structured directory for images, organizing them into labeled folders for better management.
- Review and correct labels associated with images to ensure accuracy and consistency.

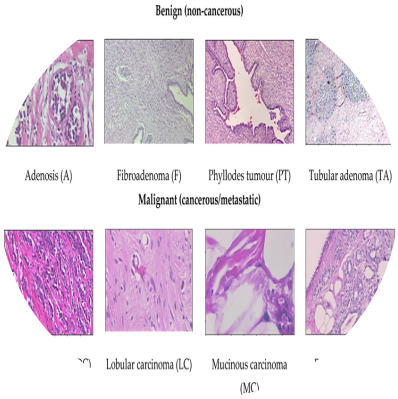
Data Description & Structure

Benign Category

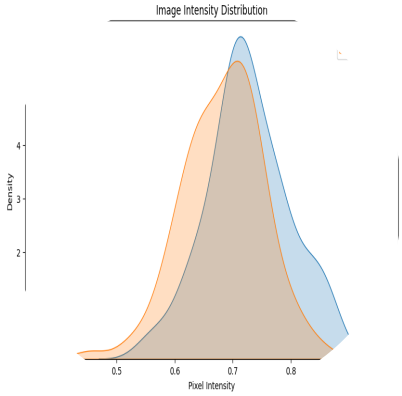
1. Adenosis: 113 - a non-cancerous condition where the breast lobules are enlarged.
2. Fibroadenoma: 260 - a common benign breast tumor made up of glandular and stromal tissue.
3. Phyllodes Tumor: 121 - are rare breast tumors that can be benign but have the potential to become malignant.
4. Tubular Adenoma: 150 - a benign breast tumor that resembles the milk ducts.

Malignant Category

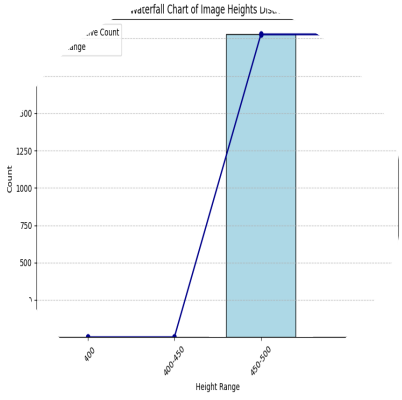
1. Ductal Carcinoma: 903 -The largest subset, ductal carcinoma refers to cancer that begins in the milk ducts and is one of the most common types of breast cancer.
2. Lobular Carcinoma: 170 - a type of cancer that begins in the lobules (milk-producing glands) of the breast.
3. Mucinous Carcinoma: 172 - a rare type of breast cancer characterized by the production of mucin.
4. Papillary Carcinoma: 142 - a rare form of breast cancer that has finger-like projections.



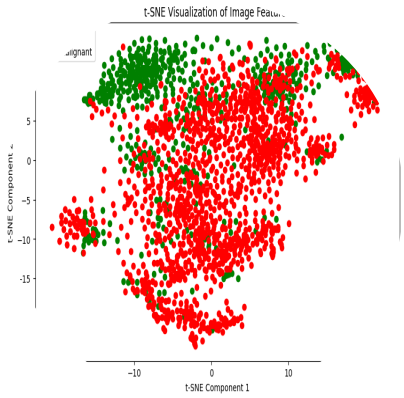
Explorative Data Analysis (EDA)



the graph suggests that while benign images tend to exhibit a more concentrated range of higher brightness values, malignant images present a wider spectrum of intensities.



The chart illustrates the distribution of image heights within the dataset. The majority of images fall within the 450-500 pixel height range, with a decreasing frequency for both smaller and larger image heights.



t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique used to visualize high-dimensional data in a lower-dimensional space while preserving the underlying structure of the data. Each point in the plot represents an image, and its position is determined by its calculated t-SNE components. Green points represent benign images, and red points represent malignant images. the t-SNE algorithm has effectively preserved the underlying structure of the data. indicating that the image features associated with these two classes are different.

Data Preprocessing

1. Resizing images to a consistent dimension is essential for feeding them into neural networks, which require fixed input sizes.
2. Normalization & Scaling by adjusting the pixel values of images to a common scale.
3. Data Augmentation artificially increase the size of the training dataset by creating variations of existing images.
4. Batching the dataset into smaller subsets that are processed together during training.

Statistical Analysis

Hypothesis Testing Hypothesis testing tested to validate assumptions about the dataset and draw inferences. It assesed the performance of different models or algorithms (comparing accuracy between a baseline model and a new model).It validated assumptions about the distribution of pixel values or class labels, ensuring that the data met the necessary conditions for modeling techniques.

Correlation Analysis Correlation analysis assessed the strength and direction of relationships between variables, feature selection and modeling strategies. Identified relationships between different features (e.g. pixel intensity values, color channels) and the target variable (class labels), guiding the selection of relevant features for the model.

MODELING

Model Architecture

- **CNN Model**: A basic convolutional neural network (CNN) architecture consisting of convolutional, pooling, and dense layers.

- **ResNet Model**: A pre-trained ResNet50 model used as the backbone, followed by global average pooling and dense layers for classification.

Model Training

Both models were trained using a combination of training and validation datasets. We used optimization algorithms to minimize the loss function and update model parameters. Hyperparameters such as learning rate, batch size, and number of epochs were tuned for optimal performance.

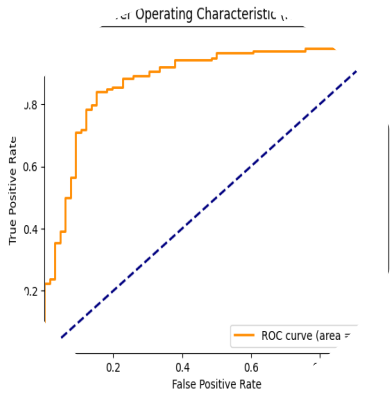
Model Evaluation

- **Accuracy**: Overall proportion of correct predictions.
- **Precision**: Proportion of positive predictions that were correct.
- **Recall (Sensitivity)**: Proportion of actual positive cases correctly identified.
- **Specificity**: Proportion of actual negative cases correctly identified (primarily for ResNet model).
- **F1-score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Detailed breakdown of correct and incorrect predictions.
-
- **ROC Curve**: Visual representation of the model's ability to distinguish between classes. These metrics provide insights into the model's ability to correctly classify malignant and benign breast cancer samples. Additionally, visualizations such as confusion matrices were created to compare the predicted labels against the actual labels.

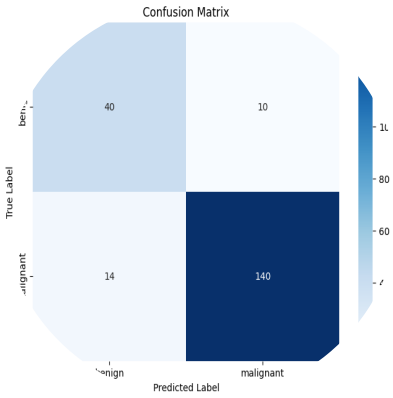
Findings and Results Interpretation

CNN Model

The CNN model achieved reasonable accuracy, but exhibited signs of overfitting, as evidenced by the gap between training and validation accuracy.



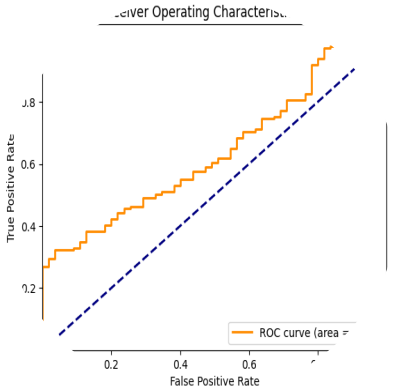
The ROC curve is closer to the top-left corner, indicating that the model's performance is good. This means a high TPR (sensitivity) with a low FPR, indicating that the model can correctly identify most positive cases with few false positives.



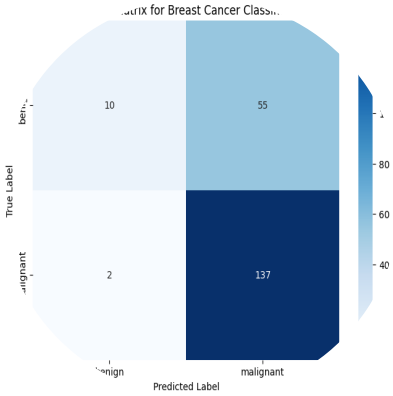
The low number of false negatives suggests that the model is effective in detecting malignant cases, which is crucial for early diagnosis. While the number of false positives is relatively low.

ResNet Model

The ResNet model demonstrated superior performance, with higher accuracy, precision, and recall. The confusion matrix revealed a lower false positive rate for the ResNet model. The ROC curve indicated a strong ability to discriminate between benign and malignant cases.



A larger AUC value indicates a better model. In this case, the AUC is 0.64, suggesting moderate discriminative power. The ROC curve illustrates the trade-off between sensitivity and specificity. As the threshold for classification changes, the TPR and FPR will vary, affecting the model's performance.



The confusion matrix indicates that the model has a strong ability to correctly classify malignant cases while maintaining a relatively low false positive rate.

CONCLUSION

This project demonstrates the potential of machine learning models in improving breast cancer diagnosis in Kenya. By accurately classifying histopathological images into benign and malignant categories, the model can significantly reduce diagnostic errors and expedite the diagnostic process. This advancement is crucial for addressing the high breast cancer mortality rates in Kenya, where late-stage detection is prevalent.

The ResNet model outperformed the CNN model in classifying breast cancer images, demonstrating higher accuracy, precision, and recall. The model's ability to effectively differentiate between benign and malignant cases holds promise for improving breast cancer diagnosis.

Limitations

- The performance of the models is dependent on the quality and quantity of the training data.
- The current study focused on binary classification (benign vs. malignant); further research is needed to explore multi-class classification for different breast cancer subtypes.
- The generalization of the models to unseen data from different sources requires further evaluation.

Key Challenges

- Limited access to quality data and computational resources.
- Addressing ethical considerations related to medical data.
- Ensuring model interpretability and explainability.

RECOMMENDATIONS

1. **Integration into Healthcare Systems**: Integrate the developed model into the Kenyan healthcare system to assist pathologists and enhance diagnostic accuracy.
2. **Training and Education**: Provide training for healthcare professionals on the use and interpretation of the machine learning model to ensure effective implementation.
3. **Data Expansion**: Continuously expand and update the dataset with new histopathological images to improve the model's robustness and accuracy over time.
4. **Ethical Considerations**: Address ethical considerations related to patient data privacy and ensure compliance with relevant regulations.
5. **Public Awareness**: Increase public awareness about the importance of early breast cancer detection and the role of advanced diagnostic technologies in improving outcomes.
6. Tracking of key metrics, detecting model drift, and implementing updates or retraining as needed to adapt to changes in data or user behavior.

Next Steps

- Expand the dataset to include a larger and more diverse set of images.
- Develop a user-friendly interface for healthcare professionals to interact with the model.
- Conduct a pilot study in a clinical setting to assess the model's impact on patient outcomes.
- Model refinement

GROUP MEMBERS

Andrew Mutuku (https://github.com/AndrewNthale)

[Amina Saidi]

[Wambui Githinji]

[Winnie Osolo]

Joseph Karumba (https://github.com/josephkarumba)

[Margaret Njenga]