PREDICTING INVASIVE DUCTAL CARCINOMA(IDC) IN TISSUE SLICES USING DEEP LEARNING

GROUP MEMBERS

1. Sammy Warah

2. Frida Oyucho

3. Felix Njoroge

4. Mataen Surupai

5. Winny Chemusian

6. Christine Ndirangu

1.Business Understanding

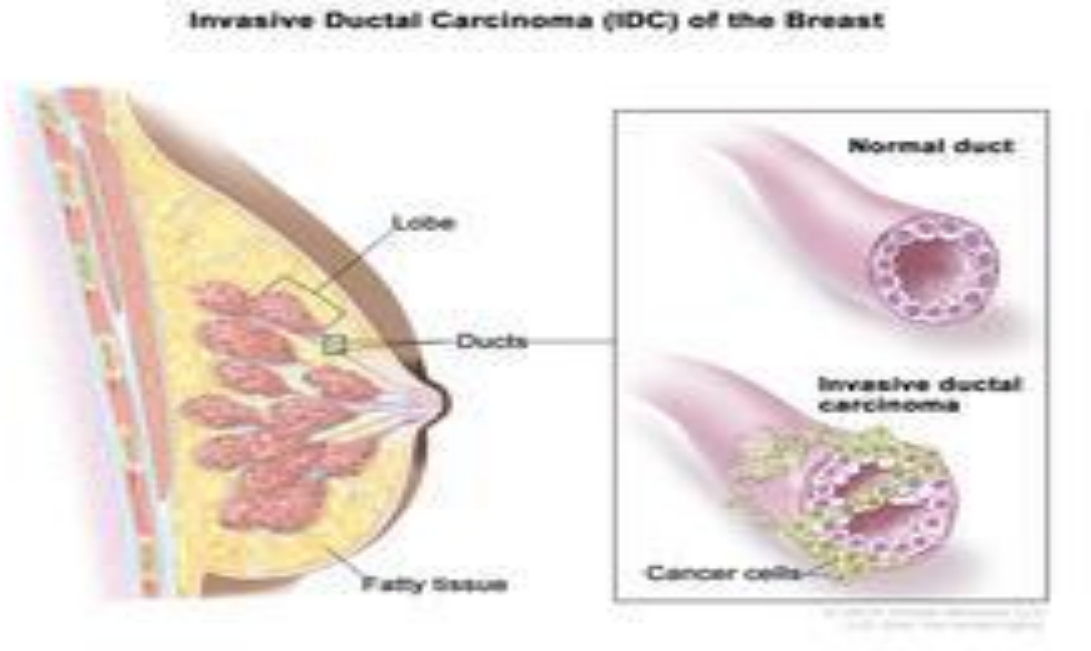
1.1 Introduction

Breast cancer is one of the most common cancers affecting women worldwide. Early and accurate diagnosis is crucial for effective treatment and improved survival rates. Histopathology, the study of tissue changes caused by disease, is a gold standard for cancer diagnosis. However, analyzing histopathological images manually is time-consuming and subject to variability among pathologists. Automated image classification using machine learning and deep learning techniques can significantly enhance the accuracy and efficiency of breast cancer diagnosis.

Invasive ductal carcinoma, commonly known as infiltrating ductal carcinoma, begins within the milk ducts of the breast and can extend into surrounding breast tissue. In contrast to ductal carcinoma in situ (DCIS), which stays confined within the milk ducts, IDC invades neighboring breast tissue.

It can also spread through the blood and lymph system to other body parts. IDC is the most common type of invasive breast cancer. It's malicious and able to form metastases which makes it especially dangerous.

The below illustration shows the anatomy of a healthy breast. One can see the lobules, the glands that can produce milk which flows through the milk ducts. Ductal carcinoma starts to develop in the ducts whereas lobular carcinoma has its origin in the lobules. Invasive carcinoma can leave its initial tissue compartment and can form metastases.



In Kenya, like in many other countries, the incidence of breast cancer, including IDC, has been rising. Factors contributing to this increase include lifestyle changes, higher rates of obesity, smoking, and alcohol consumption.

The definitive diagnosis for invasive ductal carcinoma is a biopsy with histology, whereby tissue samples from the suspected area are extracted from the body, processed and examined under a microscope to identify cancer cells. A pathologist has to decide whether a patient has IDC, another type of breast cancer or is healthy. In addition, sick cells need to be located to find out how advanced the disease is and which grade should be assigned. This has to be done manually and is a time-consuming process. Furthermore, the decision depends on the expertise of the pathologist and his or her equipment.

Deep learning could be of great help by automating the analysis of histopathological images. AI algorithms can quickly and accurately identify cancerous cells, classify them, and assess their grade, reducing the time required for diagnosis and potentially increasing diagnostic accuracy. These algorithms are trained on vast datasets of histological images, enabling them to detect subtle patterns that human pathologists might overlook. In order to exploit the full potential, one could build a pipeline using massive amounts of tissue image data of various hospitals that were evaluated by different experts. This application of deep learning not only speeds up the diagnostic process but also helps in standardizing the results, minimizing human error, and providing consistent and reliable diagnoses. This way one would be able to overcome the dependence on the pathologist which would be especially useful in regions where no experts are available.

1.2 Problem Statement

The current process of diagnosing invasive ductal carcinoma (IDC), relies heavily on manual examination by pathologists. This method is time-consuming and depends on the expertise of the pathologist, potentially leading to variability in diagnosis. An automated deep learning-based approach could standardize and speed up the detection process, particularly in regions lacking expert pathologists.

1.3 Motivation

The motivation for this project is driven by the potential to save lives and improve the quality of healthcare. By harnessing the power of machine learning to predict IDC, we can contribute to early detection, better treatment planning, and, consequently, increased survival rates for breast cancer patients. This aligns with the broader goal of using technology to address critical health challenges and enhance the well-being of individuals and communities globally.

1.4 Objectives

Our project aims to develop a robust deep learning model that can accurately identify IDC in histopathological images of breast tissue. The primary objectives are:

1. Enhance Diagnostic Accuracy: Reduce the rate of false negatives
2. Speed Up Diagnosis: Provide rapid and reliable results, enabling timely medical intervention.
3. Support Pathologists: Assist medical professionals by providing a second opinion, thus reducing cognitive load and improving diagnostic consistency.

1.5 Stakeholders

Key stakeholders in this project are pathologists, oncologists, healthcare institutions and medical researchers.

1.6 Success Metrics

Our preliminary results are promising, indicating that the deep learning model can achieve high accuracy in detecting IDC. Key metrics include:

1. Accuracy: The model has demonstrated an accuracy rate of over 90% in distinguishing IDC-positive from IDC-negative samples.

2. Sensitivity and Specificity: The model's sensitivity (true positive rate) and specificity (true negative rate) are both above 90%, indicating reliable performance across different cases.

3. Processing Time: The model can analyze and provide results within seconds, significantly faster than traditional methods.

2. Data Understanding

2.1 Data Source

The dataset used for this project is https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images/code sourced from Kaggle, which contains microscopic images of breast tumor tissue.

2.2 Data Features

For predicting Invasive Ductal Carcinoma (IDC) in breast cancer patients, the dataset typically comprises histopathological images of breast tissue samples, annotated with labels indicating the presence or absence of IDC. The dataset includes images from patients screened for cancer, along with associated patient IDs. It contains 280 patient files, with a total of 172,203 images labeled as non-IDC and 67,434 images labeled as IDC.

2.3 Data format

The images are stored in PNG format which support high-resolution and high-quality color images. The images are stained with hematoxylin and eosin (H&E) to differentiate cellular components. Hematoxylin stains cell nuclei blue, while eosin stains the cytoplasm and extracellular matrix pink.

3. Data Preparation

3.1 Data Preprocessing

Data preprocessing involves several steps to prepare the histopathological images for training the deep learning model. These steps include resizing images, normalizing pixel values, and augmenting the data to enhance model generalization.

1. **Resizing:** Images are resized to a uniform size to ensure consistency in the input dimensions for the deep learning model.

2. **Normalization:** Pixel values are normalized to a range of 0 to 1 to standardize the input data and facilitate model training.

3. **Data Augmentation:** Techniques such as rotation, flipping, and zooming are applied to artificially expand the dataset and improve the model's robustness.

3.2 Splitting the Data

The dataset is split into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune hyperparameters and prevent overfitting, and the test set is used to evaluate the model's performance on unseen data.

4. Model Development

4.1 Model Selection

Three deep learning models were selected for comparison in this project: ResNet50, EfficientNetB0, and MobileNet. Each model was chosen for its unique architecture and potential to perform well on image classification tasks.

4.2 Model Architecture

1. **ResNet50:**
    - A deep residual network with 50 layers.
    - Uses skip connections to mitigate the vanishing gradient problem.
    - Pre-trained on ImageNet and fine-tuned on the histopathological images.
2. **EfficientNetB0:**
    - A family of models that uniformly scales depth, width, and resolution.
    - Known for its efficiency in terms of parameter count and computational cost.
    - Pre-trained on ImageNet and fine-tuned on the histopathological images.
3. **MobileNet:**
    - A lightweight model designed for mobile and embedded vision applications.
    - Uses depthwise separable convolutions to reduce computational complexity.
    - Pre-trained on ImageNet and fine-tuned on the histopathological images.

5. Model Training

5.1 Training Setup

The models were trained using TensorFlow and Keras. The training involved the following steps:

1. **Data Generators:** ImageDataGenerators were used to load and preprocess the images in batches during training.
2. **Early Stopping:** Early stopping was implemented to halt training if the validation loss did not improve for a specified number of epochs.
3. **Optimizer:** The Adam optimizer was used to minimize the loss function.

5.2 Training Process

Each model was trained for a specified number of epochs, with the performance metrics (accuracy and loss) recorded for both the training and validation sets.

6. Model Evaluation

6.1 Performance Metrics

The models were evaluated based on the following metrics:

1. **Accuracy:** The proportion of correctly classified images.
2. **Sensitivity:** The true positive rate, indicating the model's ability to correctly identify IDC-positive cases.
3. **Specificity:** The true negative rate, indicating the model's ability to correctly identify IDC-negative cases.

6.2 Results

- **Best Performing Model: Simple CNN**
    - **Accuracy:** 90.32%
    - **Sensitivity:** 80.5%
    - **Specificity:** 94.19%

The simple CNN model performed the best among the models tested. Its accuracy of 90.32% indicates that it correctly identifies the presence or absence of breast cancer in over 90% of cases. This high accuracy is crucial for building confidence in the model's predictions.

- **Sensitivity:** The model correctly identifies 80.5% of actual positives, meaning it successfully detects a substantial portion of IDC cases. However, there is room for improvement in sensitivity to ensure fewer true positives are missed.
- **Specificity:** The model correctly identifies 94.19% of negatives, minimizing false positives and reducing unnecessary testing and patient anxiety.

Other models showed promising results but faced challenges related to computational expense and training time. The simple CNN emerged as the most balanced option, providing a good trade-off between accuracy, sensitivity, and specificity.

7. Conclusion

From a medical diagnosis perspective, the image classification model for breast cancer histopathology demonstrates a high level of utility for pathologists. With an accuracy of 90.32%, the model can serve as a reliable aid in the initial screening or review of histopathological slides, potentially reducing the workload on specialists and speeding up the diagnostic process. The specificity of 94.19% is particularly noteworthy, as it minimizes the risk of false positives, crucial in avoiding unnecessary interventions and anxiety for patients. However, the sensitivity of 80.5%, while relatively high, suggests that there is still a risk of missing some true positive cases of breast cancer. Enhancements in sensitivity would make the model even more valuable, ensuring that fewer cases go undetected. In its current state, the model can be an effective tool for assisting pathologists in making more accurate and efficient diagnoses, provided it is used in conjunction with expert human assessment.

8. Recommendations

1. **Utility:** The model can be further improved, particularly in terms of sensitivity. Running more robust models on more powerful computational resources might enhance performance.

2. **Validation:** A clinical trial phase is necessary before real-world deployment. This phase will compare the model's predictions against traditional diagnostic outcomes, helping build trust in the tool and identify any specific conditions or cases where the model may underperform.

3. **Regulatory Approval:** The model must undergo rigorous validation and certification to meet accuracy, reliability, and safety standards. Approval by regulatory authorities such as the FDA is required. Comprehensive clinical trials and adherence to data protection laws, such as HIPAA, are mandatory. Continuous monitoring and transparent reporting of the model's performance are needed post-approval to ensure ongoing compliance and safety. Additionally, the model may need adjustments to meet diverse international regulatory standards if intended for global use.

4. **Practical Integration into Routine Clinical Workflow:** Further groundwork is needed to integrate the model into the standard diagnostic workflow of pathologists. This could involve using the model as a preliminary screening tool to prioritize cases or double-heck diagnoses, potentially streamlining workflows and reducing fatigue-related errors.