Breast Cancer Detection Classification Analysis

Aquincum Labs

Isaac Klar - Mingi Kang - Karam Al-Askar - Shaamil Karim

Data

We used the Breast Histopathology Images dataset from Kaggle, containing 277,524 breast cancer images scanned at 40x magnification, each 50 x 50 pixels in PNG format. Of these, 198,738 are IDC-negative, and 78,786 are IDC-positive. Images follow the naming convention “u_xX_yXclassC.png” (where u is the patient ID, X and Y are coordinates, and C indicates IDC status).

Originally sourced from gleason.case.edu, the dataset (around 3 GB) was downloaded locally via the Kaggle API with the kagglehub library. A preliminary image review confirmed their quality, providing ample high-quality data for our classification analysis.

Initial Data Analysis

With many more IDC-negative images (198,738) than IDC-positive (78,786), a CNN might lean towards the majority class, potentially affecting IDC-positive classification accuracy. To address this, we plan to counter the imbalance by assigning higher weights to IDC-positive samples in the loss function, using the class weight ratio (198,738 / 78,786). Alternatively, we may use data augmentation to increase positives or reduce negatives.

Website

The Medscape article covers methods for determining breast cancer severity, focusing on invasive ductal carcinoma (IDC). It describes a scoring system based on gland formation,

nuclear atypia, and mitosis, with the Van Nuys prognostic index rating recurrence risk. While this differs from our project, the background and key features discussed are helpful.

PubMed Central provides useful images of different breast cancer types and explains indicators for each. This source also supplied images for the Kaggle dataset and reviews methods from other IDC classification models. It notes that IDC can grow in sheets, nests, cords, or as single cells.

https://emedicine.medscape.com/article/1954658-overview?form=fpf

https://pmc.ncbi.nlm.nih.gov/articles/PMC8642363/#Abs1

https://pubmed.ncbi.nlm.nih.gov/27563488/

https://pmc.ncbi.nlm.nih.gov/articles/PMC8582388/#:~:text=Microscopically%2C%20there%20are%20a%20wide,barely%20detectable%2C%20or%20altogether%20absent.

https://andrewjanowczyk.com/use-case-6-invasive-ductal-carcinoma-idc-segmentation/

https://andrewjanowczyk.com/revised-deep-learning-approach-using-matlab-caffe-python/

Next Steps

After initial data analysis, we will split the data into training and testing sets. The PNG images will be converted into Numpy 1D arrays by flattening or into tensors to retain their 2D shape. We will apply K-Fold cross-validation with a K value between 5 and 10 to balance bias and variance, and experiment with different K values. This process can be parallelized for efficiency.

For our binary classification task, we will use TensorFlow to build a CNN model with a linear output layer. The architecture will include CNN modules, Max-Pooling, Batch Normalization, Dropout, and ReLU activation layers. The Adam optimizer will update weights, and Binary Cross-Entropy Loss will evaluate performance.

We plan to train the model for 5-10 epochs and use early stopping to prevent overfitting. During training, we will track loss, accuracy, and evaluate the final test accuracy.