**Internship Project Report**

**Breast Cancer Histology Image Classification Using DenseNet**

1. **Introduction**

Breast cancer has emerged as a significant public health concern due to its highest morbidity rate among all cancers worldwide. The most prevalent subtype of breast cancer is invasive ductal carcinoma (IDC) [5]. Early diagnosis in breast cancer can increase the chances of successful treatment and survival [1][3][4]. When grading the aggressiveness of a whole mount sample, pathologists frequently concentrate on the areas that contain the IDC. As a result, identifying the precise IDC regions inside of a whole mount slide is one of the typical pre-processing processes for automatic aggressiveness rating. Pathologists can take a significant amount of time to evaluate whole mount slides for a given patient. Computer-aided Systems for diagnosis help this process run more efficiently and more inexpensively. Traditional classification strategies rely on feature extraction techniques created for a particular problem based on domain expertise. Deep learning techniques have proven to be a key alternative to feature-based approaches for overcoming their numerous drawbacks [1].

        A fundamental and crucial computer vision task is image classification. Convolutional neural networks (CNNs) have taken over as the main machine learning strategy for computer vision recognition in recent years [5]. The original LeNet5 had 5 layers, VGG had 19 layers, and Residual Networks (ResNet) have crossed the 100-layer threshold. Training these models present challenges such as too many parameters, gradient vanishing, and complicated training regimes to prevent overfitting. In comparison to models like VGG and Resnet, Dense Convolutional Network (DenseNet) has dense connection and is superior to other models. Direct connections from any layer to all subsequent layers distinguish the DenseNet model from other CNNs and potentially enhance the information. The feature maps from the previous layer are used to generate new feature maps in the following layer. Each feature map in this second layer is a combination of the feature maps in the previous layer. And the value of the feature map in the second layer is determined at each given pixel by multiplying each feature in the first layer by a convolution kernel, with a distinct kernel for each feature map in the first layer. The responses are then added to a bias term and modified using a simple non-linear technique.

In this work we have used applied densenet architecture to learn features using patches created from whole mount slide images to then use these features to classify IDC vs non-IDC [5][2].

1. **Dataset**

The original dataset consisted of 162 whole mount slide images of Breast Cancer specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch’s file name is of the format: u*xX*yY*classC.png — > example 10253*idx5*x1351*y1101*class0.png . Where u is the patient ID (10253*idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC. The images in the dataset are small image patches extracted from larger digital pathology scans. Small image patches extracted from larger digital pathology scans are used with DenseNet Block to detect IDC cancer [5]. The dataset can be found at the following link for download - <https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images>

1. **Methods**

In contrast to models like VGG and Resnet , Dense Convolutional Network (DenseNet) has dense connectivity. DenseNet can reduce the number of parameters, improve feature map propagation, and solve the vanishing-gradient problem. The images in the dataset are resized into 50 x 50 images and have exactly 3 input channels(RGB).

* 1. **Pre-processing**

We divide the data into 3 parts: training data, testing data and validation data. Data is divided into 3 sub folders: train\_seg, test\_seg and val\_seg. Each of these 3 sub-folders contain 2 folders idc-minus and idc-plus respectively.

The data was split as follows:

1. Train Data Size= 0.7\*Dataset size (1,94,266)
2. Validation Data Size= 0.21\* Dataset size (58, 281)
3. Test Data Size= 0.09 \* Dataset size (24, 977)
   1. **Data Augmentation**

Medical image datasets suffer from low sample size due to various reasons. Data augmentation is a technique by which the number of training examples can be increased to compensate for low sample size. Data augmentation in image datasets involve adding random rotations, zoom, mirroring, and noise addition to existing images in the dataset to increase the sample size. This serves two purposes, first, it increases the size of the dataset, second, it increases the robustness of our model [5]. In practical cases, the model might not always see perfectly aligned images hence these images make our model to look for signals in the image regardless of the orientation. We performed data augmentation to make the model more robust using the ImageDataGenerator class in keras to increase the size of the dataset. Each Image is rotated through 45 degree angles. The base model used in this project is the DenseNet121. The architecture of the model is shown in Fig 1.

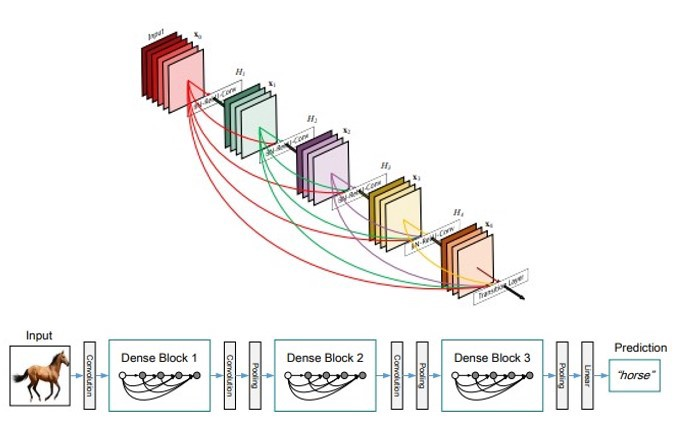


Fig1: Densenet Model

We define a sequential model. After adding the base model, we add a flattening layer to convert all the resultant 2-Dimensional arrays from pooled feature maps into a single long continuous linear feature vector of size 1024. Batch Normalization is applied to normalize output of previous layers. We then use dropout to drop random connections to gradually reduce the feature map size to 128 from 1024. This reduces the model complexity as dropout is known to improve model training time and reduce model complexity without compromising predictive ability of the model.

Table

Description automatically generated

Fig 2:  Architecture of the model used

* 1. **Training**

Due to the large number of parameters to be trained (6, 372, 226), and GPU constraints, we opted to train our model using a transfer learning approach. Briefly, transfer learning is a technique where rather than starting from random initialization of weights of a deep neural network, we set the weights using a pre-trained model trained on an image dataset or a similar dataset if available. This has been shown to heavily reduce training time, improve robustness, and can be run on single GPU machines. We imported the weights file of a densenet121 model pre-trained on ImageNet provided in the keras deep learning library. This greatly reduced our training time compared to random initialization of weights. The model was trained on cloud GPU provided by Kaggle, a google owned platform which hosts machine learning competitions. Kaggle provides 36 hours of GPU time and TPU (Tensor Processing Unit) support for 12 hours free of cost. During training, the model consumed around 6 GB of RAM and 6.1 GB of Disk Space. The GPU memory utilization is around 8.4 GB which is decent when compared to large language models and other huge deep learning model.The approximate training hour taken by the model is 1 hour.

**4) Results**

The results from the training are as follows. The accuracy and loss of the model is plotted in Fig 3 after training it for 30 epochs.

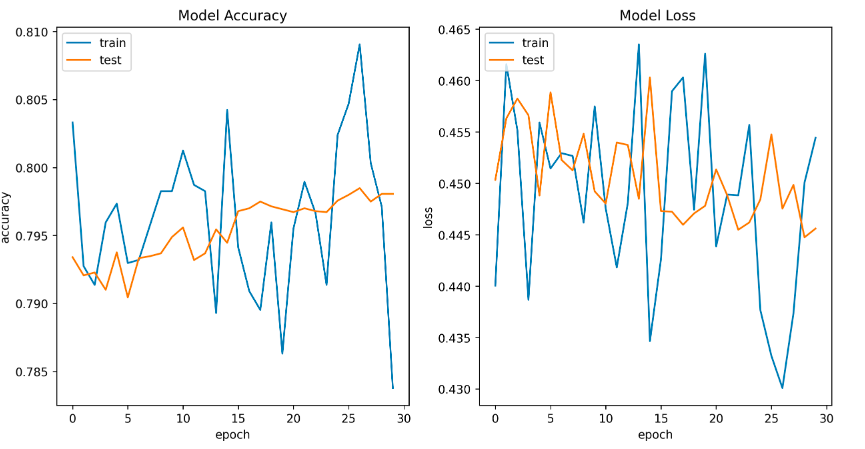


Fig 3: Report of test data is shown in Fig 4.

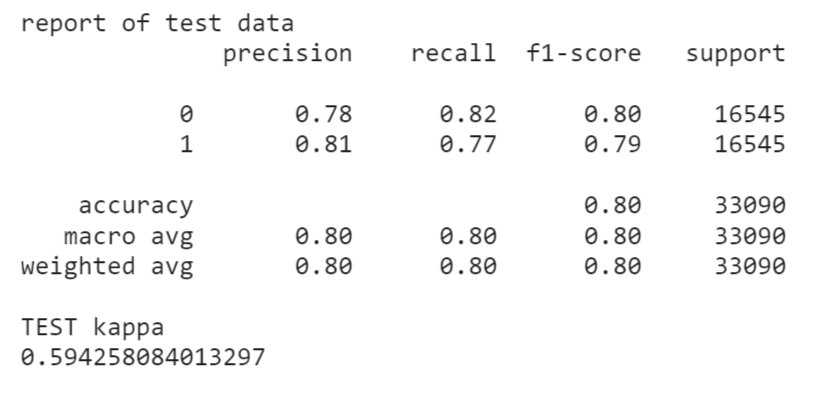


                    Fig 4: Precision, Recall, and F1 score metrics evaluated on test data

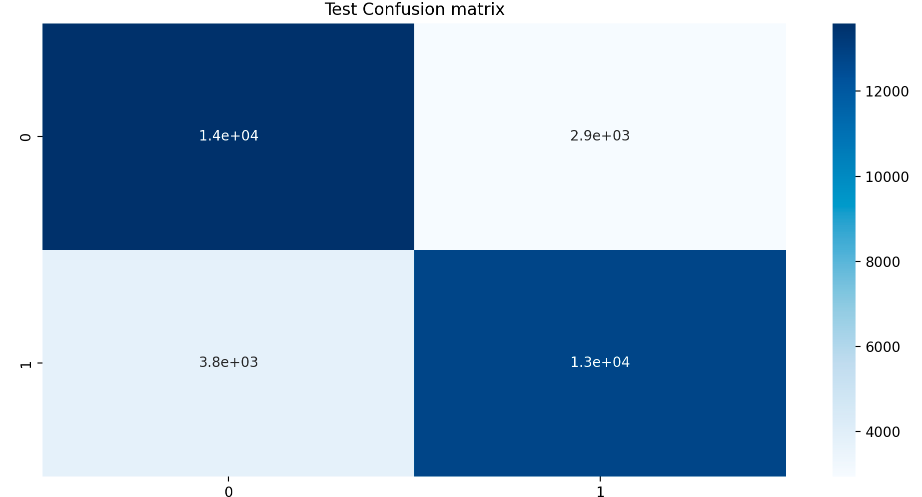


Fig 5: Confusion matrix evaluated on test data

The confusion matrix for the test data is as shown in Fig 5. The model classifies 27,000 samples out of 33,700 samples correctly. Hence, we were successfully able to train and test our deep learning model based on DenseNet architecture to classify breast cancer histopathology. We used transfer learning along with data augmentation to add robustness to our model.

**References**

1. Gupta, Vibha, and Arnav Bhavsar. "Sequential modeling of deep features for breast cancer histopathological image classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018
2. Jiménez Gaona, Yuliana, et al. "Densenet for breast tumor classification in mammographic images." *International Conference on Bioengineering and Biomedical Signal and Image Processing*. Springer, Cham, 2021
3. Aresta, Guilherme, et al. "Bach: Grand challenge on breast cancer histology images." *Medical image analysis* 56 (2019): 122-139
4. Ferreira, Carlos A., et al. "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2." *International conference image analysis and recognition*. Springer, Cham, 2018
5. H. Chapala and B. Sujatha, "ResNet: Detection of Invasive Ductal Carcinoma in Breast Histopathology Images Using Deep Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 60-67, doi: 10.1109/ICESC48915.2020.9155805.