Project 4:

Breast Cancer Detection

Convolution Neural Network with Tensorflow & Keras

Trent McNabb & Scott Stimpson

October 19, 2021

Introduction to Breast Cancer Detection

Invasive Ductal Carcinoma is the most common subtype of all breast cancers and affects the milk duct cells

Pathologist focus on the images which include IDC whether benign or malignant to determine the aggressiveness of the cancer (risk of spreading)

Our objective is to use the histology images to predict whether IDC is benign or malignant using a Convolutional Neural Network in Tensorflow & Keras and understand how it could be used to improve treatment

Data Acquisition & Exploration

Dataset was downloaded from Kaggle.com - IDC Dataset

- Consists of 162 whole mount slide images of breast cancer histology 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 patient has both non-IDC and IDC images which are separated into sub-folders (0, 1)

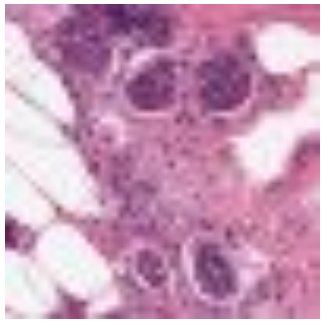
File name format: u*x*X*y*Y*classC.png* — > *example -- 10253*idx5 *x1351* y1101 *class 0.png*

*'u' is the 'Patient ID' (10253*idx5)

'X' is the x-coordinate of where this patch was cropped

'Y' is the y-coordinate of where this patch was cropped

C indicates the class where 0 is non-IDC and 1 is IDC



Data Cleaning & Preprocessing

Using the files names;

1. Split the images into “Training” (80%) and “Testing” (20%) datasets,
2. Label each image data as 0: benign or, 1: malignant
3. Using Numpy - transform images into array
4. Normalize the array by dividing by 255

Model Creation, Training & Testing

Training:

- Sequential Class Model
- 3 layers of Conv2D & MaxPooling2D
- Activation Relu
- Flatten 3D features into 1D feature vectors
- Activation changed to Softmax for the output

Testing:

- Binary Cross Entropy vs Sparse Categorical Cross Entropy
- ‘Adam’ optimizer used

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 48, 48, 32)        896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 24, 24, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 22, 22, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 11, 11, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 9, 9, 128)         73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 4, 4, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 64)                131136
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 130
=================================================================
Total params: 224,514
Trainable params: 224,514
Non-trainable params: 0
_________________________________________________________________
```

Model Development

1. Using over or under sampling to try and balance the dataset between non-IDC and IDC

2. Use scikit-learn library for further preprocessing of the images as arrays

3. Weighting the results towards false positives due to the application of the model in cancer treatment

Summary

Breast cancer is the most common form of cancer in women, and invasive ductal carcinoma (IDC) is the most common form of breast cancer.

Accurately identifying and categorizing breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce error.

The accuracy would likely be increased through balancing the dataset and implementing further preprocessing of the images as arrays