Capstone Proposal : (Breast cancer Histology images classification)

1.Domain Background

The project is a competition in site called "Grand-Challenge", this challenge was under ICIAR2018 conference it finished at January 22, 2018 February ,but code submission and ranking is still available ,The challenge is about breast cancer which is one of the leading cancer-related death causes worldwide, especially on women. However, early diagnosis increases treatment success. Proper analysis of histology images is essential. Specifically, during the diagnosis procedure, specialists evaluate both overall and local tissue organization via whole-slide and microscopy images. However, the large amount of data and complexity of the images makes this task time consuming and non-trivial. Because of this, the development of automatic detection and diagnosis tools is challenging but also essential for the field.
As this is a challenge there isn't academic research available for publicly, But I think that this is appropriate Breast cancer histopathological,This paper could be relevant as it uses the same method I will use which is CNN .

---

2. Problem Statement

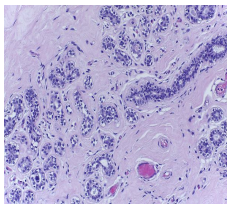
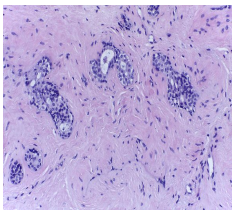
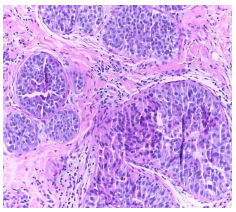
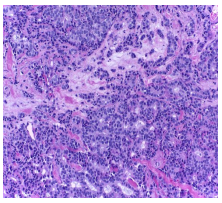
The problem is with histology images is too difficult even if on the expertise of the field because it contains a lot of information , Doctors in breast cancer can easily detect and analysis mammogram images , But in histology images it's not easy there is a tiny tumors that can't be seen or doctors ignore and it could be the key of the problem . to summarize we can ask our self
What problem exactly are we trying to solve? what is being predicted? and what is the input?
The problem is to correctly classify the type of image if it's normal or benign or else .
We predict the type of cancer from the available 4 types .
The input is 400 images labeled each type of cancer have 100 images .

---

3. Dataset

The dataset consists of microscopy images available in ".tiff " format. which are labelled as (normal, benign, in situ carcinoma or invasive carcinoma) according to the predominant cancer type in each image. The annotation was performed by two medical experts and images where there was disagreement were discarded,The dataset contains a total of 400 microscopy images, distributed as the following:

Normal 100 | Benign 100 | in situ carcinoma 100 | invasive carcinoma 100

The images are represented each type has it's label respectively,This is sample images of each type,The data size is nearly "20 GB" so it's very big to upload so I put samples of it in folder called "dataset sample" supplied in the zip , These images would be the input of our CNN after preprocessing of it. to summarize we can ask our self

What are the dimensions are the images? are they all of the same dimensions? how many images are present per class? is there a class imbalance?

The images are (2048*1536).
They all the same size .
100 images per each class and number of classes is 4.
No there isn't class imbalance

---

4. Solution Statement

We would use CNN to train our model with this dataset by using of cnn we could see tiny cells and correctly diagnose it .our model would train in all 400 image we will split our data to train and test than supply this images to the model which could recognize the pattern of each cancer type , the solution is quantifiable as we can represent our images in arrays and pass it through cnn filter which could find the best math formula and weights to get the best results. Also the solution is replicable as we increase our dataset we would get more and more accurate results the same to the number of epochs .
Diagnosing cancer type could help Save the life of millions as the early stages could be treated with very big survival rate near to 99 % .

---

5. Benchmark Model

The results that I would compare my model with it is competition results which is supplied in this link results . the highest rank is for "Sai Saketh Chennamsetty" he get accuracy of 87 % .also the main benchmark is described as following it could accessed here Breast cancer histopathologica

A description would need to be provided for the chosen benchmark? What model is it? and what is it made of?

The benchmark is Breast cancer histopathological image classification .

It’s a medical model based on the same idea of my project to classify the types of cancer . the model is using convolutional neural networks with small SE-ResNet module

---

6. Evaluation Metrics

Actually I am recommending working in (Fb score) as this is a medical project and sensitivity and specificity is very important , But I would work with the same metrics as the challenge work which is the accuracy which considers the correct classification of the microscopy images based on the overall prediction accuracy, i.e., the ratio between correct samples and the total number of evaluated images.
For accuracy to be considered appropriate, some discussion should be provided on the distribution of the target class. Is there a class imbalance?
There is claa balance so i think we can use accuracy and there is no need for any other metrics . and I should use it to compare my results with competition results which was based on accuracy .

---

7. Project Design

Our images is good to train on it as it filtered by doctors , Firstly we will use this command to download the data to our colab “!wget link of the data “ . After the download we would have “zip file” we then use “zipfile” library that help us to unzip our data .
Now we have 4 folders contain each type of cancer each one has it’s laped type , so the first thing is to upload this data and label it we will use one hot encoding so if the image was for normal the label would be [1 0 0 0] we put one in the first element to represent the type .

This is some of the libraries that we would need :
 numpy ,matplotlib , os, cv2, tqdm,tensorflow ,keras.preprocessing.image, Sequential, and keras.layers , All of it is available and for free .

We convert our images to array ,then normalize it by dividing by 255 , then translate it to grayscale and then we resize our images using cv2.resize to reduce the number of pixels on it and make it easy for the model to train and also to reduce the time and cost of training .
We then split our data to train and test and use random.shuffle to shuffle our training data .
Then after finishing data preprocessing we start of the structure of our cnn by putting the layers of it ,We would use relu activation and MaxPooling2D to reduce the time needed to train . after that our model would be able to accept any image as input and then correctly classify it .

to summarize we can ask our self :

How will the data be split? what validation methods will be considered?

A more detailed description of the CNN model's architecture should be provided. Will transfer learning be considered?

We will useK-Folds Cross Validation In K-Folds Cross Validation we split our data into k different subsets (or folds). We use k-1 subsets to train our data and leave the last subset (or the last fold) as test data. We then average the model against each of the folds and then finalize our model. After that we test it against the test set.

The CNN model contains a high capacity that can represent various functions while not requiring extracting features manually. Therefore, we use CNN to automatically extract the characteristics of breast cancer images and take full advantage of them for classification. We design a novel CNN architecture for the classification of breast cancer images using VGG-19 that mean that we would use transfer learning as it would improve our accuracy a lot .
We used Max Pooling which performs a lot better than Average Pooling.it consist of A convolutional layer extracts features from a source image.A pooling layer that downsamples each feature to reduce its dimensionality and focus on the most important elements.
A fully connected layer that flattens the features identified in the previous layers into a vector, and predicts probabilities that the image belongs to each one of several possible labels.
This is examples

```
model.add(Conv2D(256, (3, 3), input_shape=X.shape[1:]))
model.add(Activation('relu'))  # adding relu to make output range from 0 to 1
model.add(MaxPooling2D(pool_size=(2, 2))) #add maximum poll that is better than average
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Flatten())  # this converts our 3D feature maps to 1D feature vectors
model.add(Dense(64))
model.add(Dense(1))
model.add(Activation('sigmoid'))# sigmoid that range from -1 to 1
```