Histological Cancer Detection

By Rucha Waghulde
Siddharth Mandgi
Pranati Kaza

Contents

Introduction

Our aim is to create an algorithm to identify metastatic cancer in small image patches taken from larger digital pathology scans.

Since ancient times, specimens have been observed on glass slides under a microscope for Pathology diagnosis.

In recent years, attempts have been made to capture with a scanner and save it as a digital image (WSI - Whole Slide Image)

Now we can analyze WSI’s and run diagnosis using digital image analysis based on machine learning algorithms.

Machine Learning uses various techniques to model the progression and treatment of cancerous conditions.

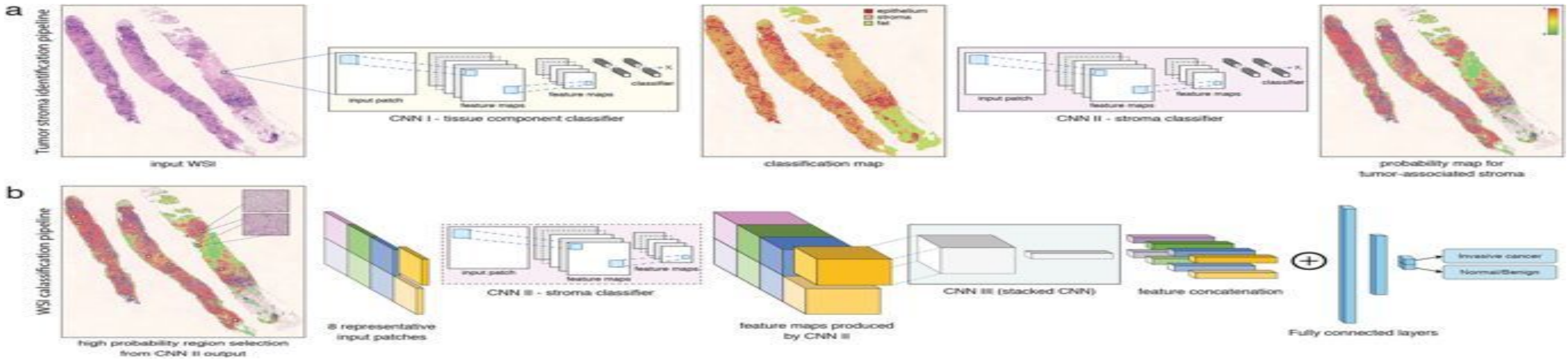
Such as, Convolutional Neural Networks (CNNs), Transfer Learning, Support Vector Machines (SVMs), Decision Trees etc.

In our project we have used CNN technique to detect and classify cancer as Malignant or Benign.

Convolutional Neural Networks

Convolutional neural networks (CNN) are deep artificial neural networks that are used primarily to classify images, cluster them by similarity and perform object recognition.

It can be used to write algorithms that can identify tumours. Hence, we are using convolutional neural networks to classify metastatic tissues as malignant or benign (1 or 0)



CNN Operations

- **Convolution**- extract features from input image. It preserves the spatial relationship between pixels by learning image features using small squares of input data.
- **Batch Normalization** - reduces internal covariate shift in neural networks. It adds an additional step between the layers.
- **Pooling** - Progressively reduces the spatial size of the representation to reduce the number of parameters and computation in the network.
- **Dropout** - it is a regularization technique. It randomly selects neurons and are ignored during training. Hence, the network becomes less sensitive to the specific weights of neurons.

These operations are the basic building blocks of every Convolution Neural Network.

Experimental Setup

We have taken many small pathology images each of 96x96px to classify.

Each file has been named with an image id

Our aim is to predict the labels of the images in the test folder.

A positive label indicates that the centre 32x32px region of a patch contains at least one pixel of tumour tissue.

Note that, the tumour tissues in the outer region of the patch does not influence the label. The outer region is provided to enable fully-convolutional models that do not use zero padding. This ensures consistent behaviour when applied to WSI.

Implementation

Provided with a set of images for training and validation datasets that contain metastatic tissues which either malignant or benign.

First code pic

Implementation

Exploratory Data Analysis (EDA) : This is carried out in three steps:

1. Loading the images for visualization

   Code

2. Data distributed into two classes - malignant and benign

Code

3. Data Augmentation

Code

Checking the distribution data

Code

Results

We were able to read and separate the images into positive and negative samples.

We created a 32x32px patch and were able to identify tumor tissue within.

Pic

Results

The accuracy of our model based on CNN is 93%

We have plotted a ROC curve with its AUC value is 0.938

We were able to predict if the level of tumor or the cancer in a metastatic tissue is malignant or benign.

Challenges Faced

- Image classification is a subject of intersection between machine learning and deep learning. It has been challenging to understand and implement these concepts.
- Certain processes in our code required high computational power. This resulted in high load on the CPU.
- Working with different kernels - Kaggle, Google Collab, Microsoft Azure, they had their own GPUs which sped up the compilation of our code.

Future Work

- The overall accuracy of the system can be improved using 3D Convolutional Neural Network and also by improving the hidden neurons with deep network.
- An improvement in classification quality could be achieved by adding clinical data (eg, age, gender, race, skin type, and anatomic location) as inputs for the classifiers. This additional information is advantageous for the decision making.
- unavailability of dataset, privacy and legal issues, dedicated medical experts, non standard data machine learning algorithms etc.

References