Collagen Fiber Extraction and Analysis in the H&E Dyed Cancer Tissue Images

Marius Latinis, marius.latinis@mf.stud.vu.lt
Mindaugas Morkūnas, mindaugas.morkunas@mif.vu.lt
Povilas Treigys, povilas.treigys@mif.vu.lt

Institude of Data Science and Digital Technologies, Vilnius University, Akademijos str. 4, Vilnius, Lithuania

December 15, 2019

Abstract

This paper investigates the use of artificial intelligence for predicting cancer patients diagnosis from collagen fibers located in the digital Hematoxylin and Eosin stained tissue images. The proposed technique uses a convolutional neural network to extract a collagen mask from the image. The collagen mask is then fed into a second neural network that predicts the diagnosis. We conclude that generating meaningful prediction solely from collagen fibers is not enough.

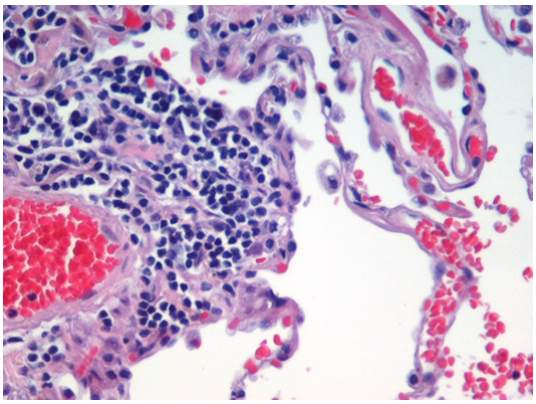
keywords: Convolutional Neural Network, Collagen Fibers.

1 Introduction

Visual examination of a tissue is a major part in cancer diagnosis. A trained pathologist is given a sample of a tissue to inspect. By looking at the tissue placed under a microscope or inspecting its digital image a pathologist diagnoses the severity of cancer and makes a prediction about the further treatment.

Unfortunatelly, manual inspection is prone to error and inter-observer variability (e.g. different pathologists are trained differently). This creates an interest to investigate whether a computer can perform the tissue inspection and provide predictions. In this project I seek to automate one aspect of tissue inspection.

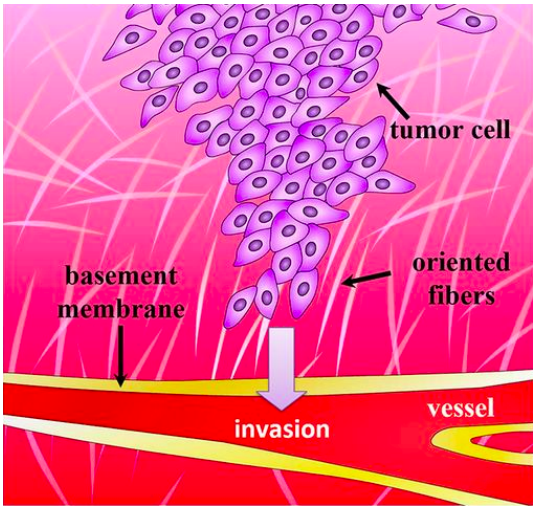
Different components of the tissue must be coloured with different colours so that the visual image of a tissue can be analysed. Hematoxylin and eosin stain (abbreviated as H&E stain) is one of the principal tissue stains used in histology:



The hematoxylin stains cell nuclei blue whereas eosin stains the extracellular matrix pink, with other structures taking on different shades, hues and combinations of these colors. The stain shows the general layout and distribution of cells and provides a general overview of a tissue sample's structure. It is the most widely used stain in medical diagnosis. When a pathologist looks at a biopsy of a suspected cancer, the histological section is likely to be stained with H&E.

Collagen is one particular component of a tissue. It is a protein located in the extracellular matrix, which becomes pink after the H&E staining. It is believed that the metastatic tumor cells interact with oriented collagen fibers to invade the blood vessels. Several studies have shown a link between collagen remodeling and the invasion and progression of mammary cancer in mouse models [1,2,3].

A specific Tumor Associated Collagen Signature 3 (TAC-3) is defined to emphasise this interaction. Image is considered to be TAC-3 positive if it contains many straight collagen fibers aligned normally to the epithelial cells boundary regions:

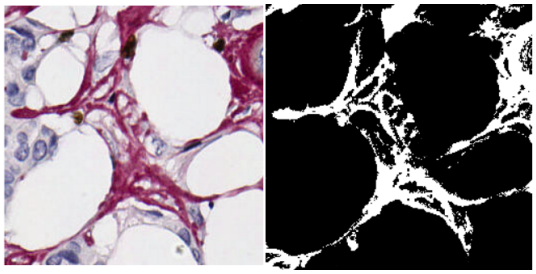


Thus, collagen fibers organisation in the tissue could be used as a biomarker to diagnose the patients cancer severity. This project aims to automate collagen as a biomarker extraction and assesment procedures.

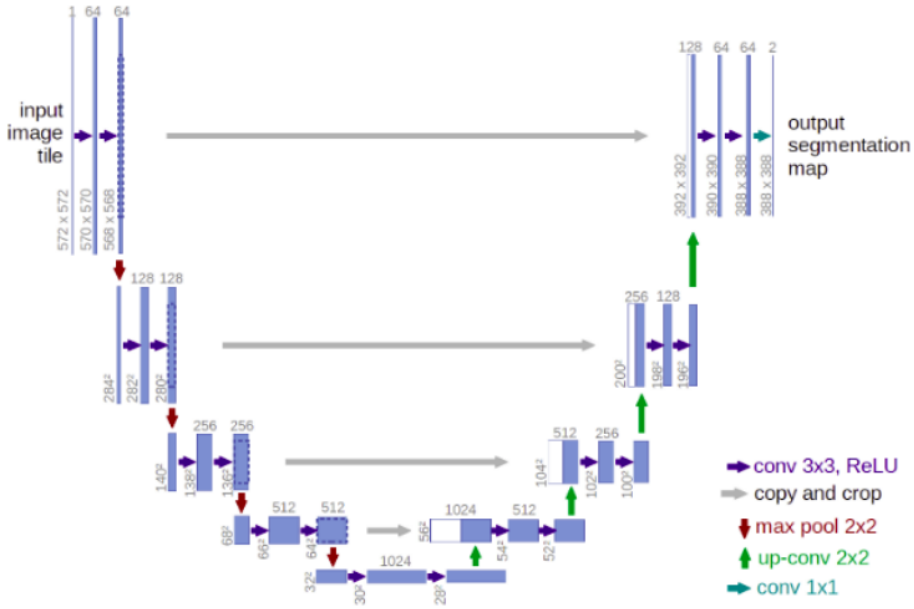
2 Collagen Extraction

The first step in this project is to take the H&E stained image containing all tissue components and extract only the shape of a collagen present in the image. This is known as the image segmentation problem. In this case a pixel representing the collagen will be labelled as a foreground, whereas any other pixel will be labelled as a background. Machine learning was used to solve this problem. A model is created and provided with the training data. The model learns to separate pixels into collagen and background classes and can be applied to unseen images.

The training data was provided by the National Center of Pathology, Vilnius. Large H&E images were tiled into $256 \cdot 256$ tiles and for each tile a collagen mask was manually extracted:

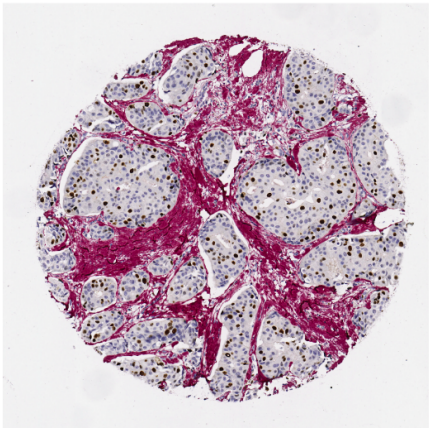


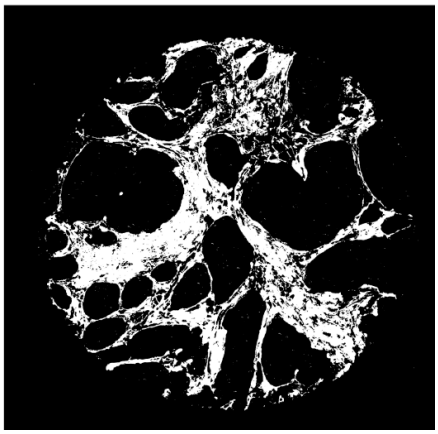
In total 43 image-mask pairs were used to train the model. A convolutional neural network U-net was chosen as a model:



The model takes a full $256 \cdot 256$ (x3 for each of the RGB colour) image as an input and performs a few downsampling procedures to infer the deep features from the image. Each downsampled layer is then upsampled and connected to the layer above to apply the inferred features on the image. Eventually the model predicts the $256 \cdot 256$ mask.

The neural network was trained on the 43 image-mask pairs using the cross-validation technique. Binary crossentropy was used as a loss function. Accuracy (i.e. proportion of correct foreground / background pixels) metric was used to test the model. All code was written in python using Keras neural network API. After the model was trained, it was applied on a new larger scale image to test its validity. The larger image was divided into $256 \cdot 256$ tiles and the model was applied to each tile to extract a collagen mask. The extracted masks were then merged into a large mask:





3 Features Generation

After the collagen is extracted, the next step is to analyse it and compute the features that can allow to diagnose the patients. More specifically, a H&E stained image is given corresponding to one patient. We wish to inspect the image and predict the cancer severity stage for that person. The following prediction pipiline is executed for this task:

- A collagen mask is computed from the image. For that the image is split into $256 \cdot 256$ tiles and the trained model is applied to each of the tile. The tiles are then merged into a single mask. Noise is removed from the mask.
- A skeletonization function is applied on the mask. This transforms a white collagen regions into 1 pixel wide representations. The skeleton tree is chopped into individual edges and each edge gets assigned a unique id.
- A breath first search algorithm is then applied to find for each collagen pixel in the mask its closest skeleton edge. In this way the collagen is divided into regions and for each region we can compute its features.
- For each region we compute several statistics. These include region length and width, as well as haralick parameters. We then use these parameters as features to predict the severity stage of cancer for each patient.

4 Output prediction

The section above gives multiple features (i.e. a list of numerical values) for each H&E stained image. The final step is to use this feature list to predict the outcome. We have a cohort of 92 patients. 71 of them are related with the positive outcome, 21 with the negative. We split the data into training (70) and test sets (22). We train a binary classification model on the training data (a

neural network with 2 dense hidden layers was chosen) and evaluate it on the test data. Unfortunatelly, the trained model does not perform better than the one which would assign outcomes randomly with a probability proportional to the sizes of the positive and negative outcome training sets.

5 References

1. Provenzano PP, Eliceiri KW, Campbell JM, Inman DR, White JG, Keely PJ. Collagen reorganization at the tumor-stromal interface facilitates local invasion. BMC Med. 2006;4:38.

2. Provenzano PP, Eliceiri KW, Yan L, Ada-Nguema A, Conklin MW, Inman DR, et al. Nonlinear optical imaging of cellular processes in breast cancer. Microsc Microanal. 2008;14:53248.

3. Provenzano PP, Inman DR, Eliceiri KW, Knittel JG, Yan L, Rueden CT, et al. Collagen density promotes mammary tumor initiation and progression. BMC Med. 2008;6:11.