Collagen-based Survival Analysis in the Hematoxylin and Eosin Stained Histological Images

Master Thesis

Systems biology master program

Vilnius university

STUDENT NAME: **Marius Latinis**

STUDENT NUMBER: **1268095**

SUPERVISOR: **Povilas Treigys, Mindaugas Morkūnas**

SUPERVISOR DECISION: **……..**

FINAL GRADE **……..**

DATE OF SUBMISSION: DD MMMM 20YY

CONTENTS

LIST OF ABBREVIATIONS

Breath First Search (BFS);

Convolutional Neural Network (CNN);

Hematoxylin and Eosin (H&E);

Red, Green, Blue (RGB);

Tumor, lymph nodes, metastasis classification (TNM)

INTRODUCTION

Visual tissue examination is a major part of a cancer diagnosis. A trained pathologist is given a tissue sample 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 further treatment.

Different tissue components must have different colours so that the visual image can be analysed. Hematoxylin and eosin (H&E) stain is one of the principal tissue stains used in histology [10]:

The hematoxylin stains cell nuclei blue whereas the eosin stains the extracellular matrix pink, with other structures taking on different shades, hues and combinations of these colours. 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 suspected cancer, the histological section is likely to be stained with H&E.

Unfortunately, manual image 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 tissue inspection and provide predictions. Furthermore, in this project, we are given a dataset of images, for which pathologists at the State Pathology Centre do not have a clear decision on how to assess them. Thus, this project aims to propose a new method that can inspect the H&E stained images automatically and generate a prediction about the patients' fate.

AIM AND TASKS

The aim: present a method that automatically analyses a digital H&E image and generates a prediction about the patient's fate.

Since the analysis is based on the extracellular protein collagen, the aim breaks into the following tasks:

- gather data: a list of patients with information (e.g. survival time) and a digital image per patient;
- write a method that extracts collagen from the image;
- compute features using the extracted collagen;
- use the computed features to make a prediction;
- evaluate the whole pipeline that makes a prediction;

LITERATURE REVIEW

This section describes the process of how a pathologist assesses a cancer tissue.

When a patient is suspected of having cancer, the disease needs to be diagnosed. Taking a biopsy is the most popular diagnosis method. A biopsy is a procedure in which the doctor removes a sample from the tissue. The tissue removed during a biopsy is cut into thin sections, placed on slides, and stained with dyes (H&E stain in the scope of this project) before it can be examined under a microscope. A pathologist looks at the tissue under a microscope to see if the tissue represents cancer. The pathologist describes the findings in a pathology report. The report contains details about the diagnosis including the cancer classification according to the TNM staging system [31]:

- The size of the primary tumour (T), measured in centimeres;
- The number of lympf nodes (N) cancer has spread into;
- Whether metastasis (M) to distant organs (beyond regional lymph nodes) has happened;
- The grade (G) of cancer cells (i.e. they are "low grade" if they appear similar to normal cells, and "high grade" if they appear poorly differentiated);

A pathology report containing the TNM staging information helps determine the best treatment option. However, the TNM system is limited. Firstly, a competent pathologist is required to stage cancer according to the standard. This makes the method costly (in comparison, this project aims to make a computer perform a diagnosis, that will become cheaper once the software is developed). Secondly, the TNM classification sometimes fails [3] to stratify patients by overall survival. Thus, we aim to propose a new method of stratifying cancer patients that could supplement the TNM classification.

This section explains the significance of protein collagen in cancerous tissue

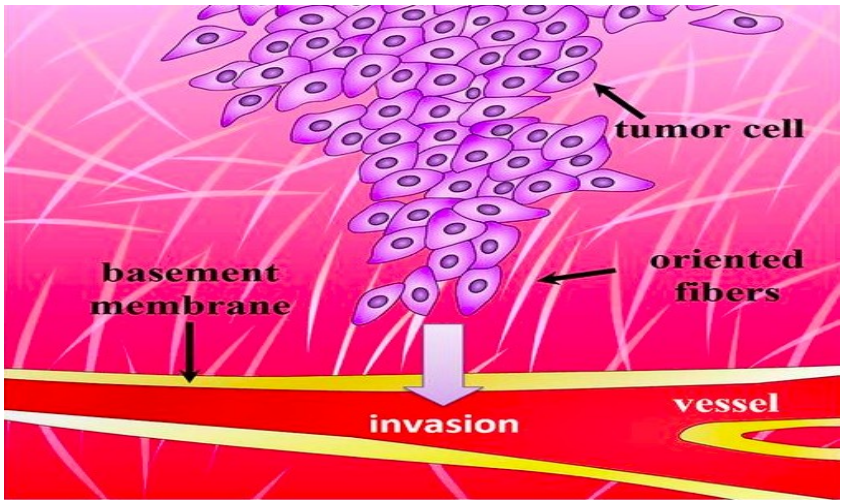
The idea of this project is to focus on protein collagen. Collagen is one particular tissue component. It is the main protein residing in the extracellular matrix. H&E stains the protein pink. It is believed that the metastatic tumour cells interact with the oriented collagen fibers to invade the blood vessels [18].

Provenzano et al. [20] first brought collagen relation with cancer into attention by analysing the tumour progression in mice. They concluded that collagen facilitates tumour progression and invasion. The so-called tumour associated collagen signature (TACS) nomenclature was introduced for describing the collagen alignment patterns:

TACS-1: characterized by dense wavy collagen bundles that serve as a reliable hallmark for locating the tumour during the early stages.

TACS-2: characterised by taut collagen fibers, because they tend to stretch as the size of the tumour increases.

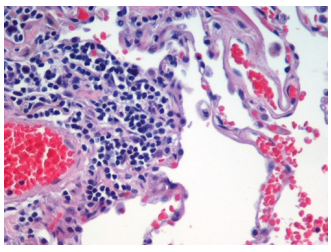
TACS-3: characterised by straight collagen fibers arranged perpendicularly to the tumour border. This rearrangement provides a pavement for the malignant cells to move along the fibers toward the stroma:



TACS screening was suggested to be a potential clinical diagnostic tool: TACS-3 collagen could indicate a poor survival rate suggesting that quantifying collagen alignment could be an independent prognostic marker. Analysing the collagen fibers in the image is the main part of this project.

This section explains the Hematoxylin and Eosin staining.

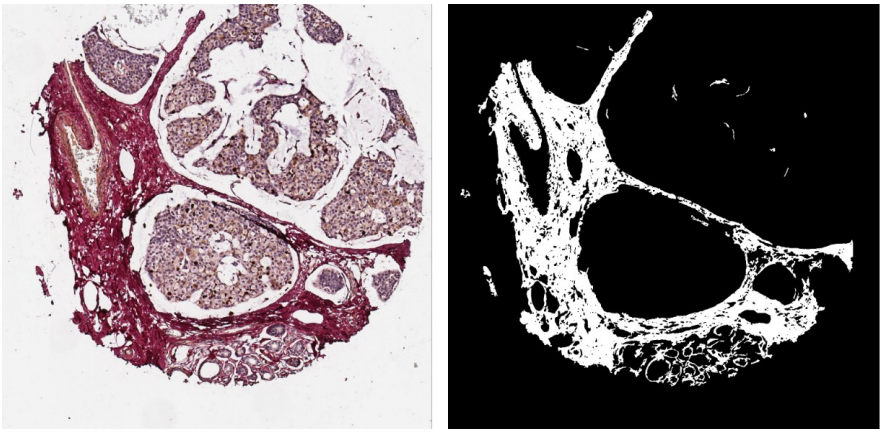
Many collagen studies were carried out on the images obtained using the Second Harmonic Generation (SHG) microscopy [29, 7]. This is an expensive and non-standard microscopy technique that is not always available in all labs. The images obtained from the National State of Pathology are instead stained with the H&E stain [10]:



The Hematoxylin stains cell nuclei blue whereas the Eosin stains the extracellular matrix pink, with other structures taking on different shades, hues and combinations of these colours. The stain shows the general layout and distribution of cells and provides a general overview of a tissue sample's structure. The H&E staining procedure is the principal stain in histology because it can be done quickly, is cheap, and stains tissues in such a way that a considerable amount of microscopic anatomy is revealed, and can be used to diagnose a wide range of histopathologic conditions. When a pathologist looks at a biopsy of a suspected cancer, the histological section is likely to be stained with H&E. Thus, due to the H&E stain popularity, the project analyses the images acquired using this stain.

This section explains the image segmentation problem and the methods for solving it.

A part of the project is to extract the collagen from H&E the picture, so that it can be used for further processing:



This is known as the image segmentation problem [13]. Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share common characteristics. In this case there will be two labels: a white label for collagen and a black label for everything else.

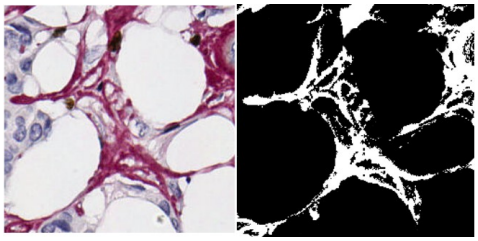
Deep learning using the Convolutional Neural Networks (CNN) is the modern approach to solving the image segmentation problem [8, 9]. AlexNet was the pioneering CNN that won the ILSVRC-2012 contest with a 84.6% test accuracy [16], while the closest competitor using the traditional techniques instead of deep architectures achieved a 73.8% accuracy. Since then new networks were introduced, each outperforming the predecessor in terms of accuracy (VGG, GoogleNet, ResNet).

The challenge to segment an H&E image is the small amount of training data available. For example, in this project, the labelled training data is generated by hand and is limited only to ~40 images. A CNN called U-net was introduced [26] to address the small training data size problem. The U-net won the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Using the same network trained on transmitted light microscopy images (phase contrast and DIC) it won the ISBI cell tracking challenge 2015. These achievements proved such architecture efficiency on medical data. Therefore, the fine-tuned version of the U-net is used in this project to solve the collagen segmentation problem.

METHODS & RESULTS

Collagen extraction

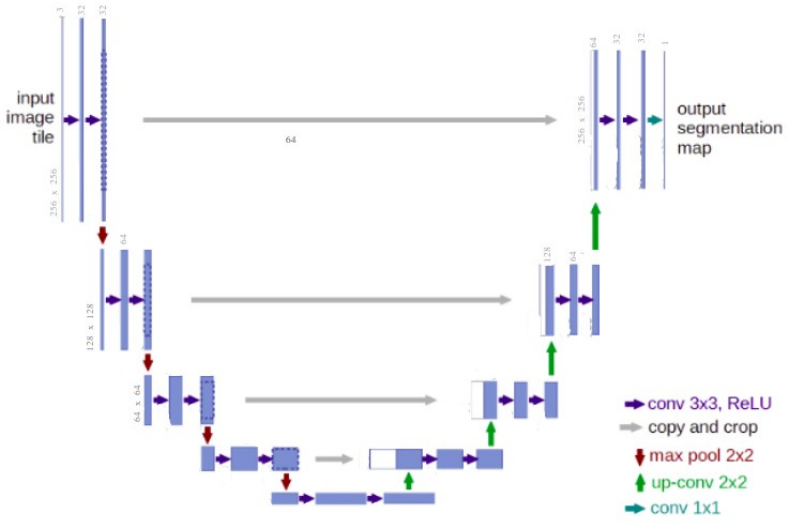
The first step in this project is to take the H&E stained image containing all tissue components and extract only the shape of a collagen present in the image. This is known as the image segmentation problem. In this case, a pixel representing the collagen will be labelled as a foreground (a bit with value 1), whereas any other pixel will be labelled as a background (a bit with value 0). Supervised machine learning was used to solve this problem. Firstly, we prepared a set of labelled samples to train the model. Each sample is a pair of two images:

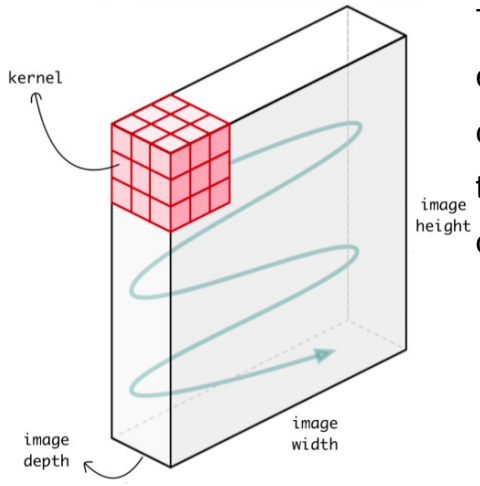


the first image is a 256 x 256 RBG tile of an H&E stained section containing collagen and other components. The second image is a collagen mask that was manually segmented using the paint software.

Each image was chosen so that the collagen in the mask would occupy not less than 20% and not more than 80% of the total area. We created a training list of 43 such samples and the segmentation correctness was validated by the pathologist expertise from the National Centre of Pathology.

Then, the convolutional neural network U-net (mentioned above) was chosen as a model for this segmentation problem. Below is the model's architecture:





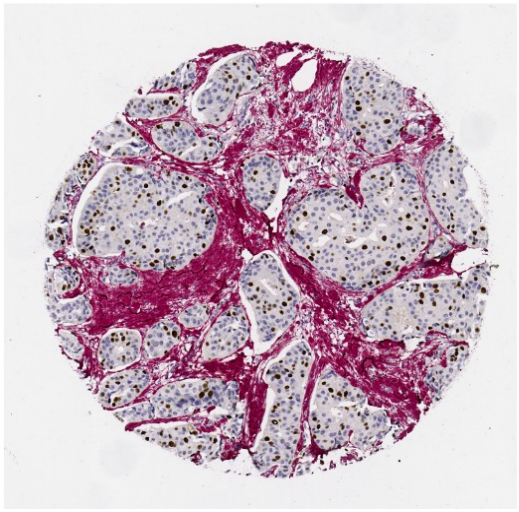
The model takes an RGB image of width and height equal to 256 (so 256 x 256 x 3). It then performs a 2D convolution operator where the kernel slides along the two dimensions (the width and the height) of the image data. Multiple kernels (32) are used in parallel.

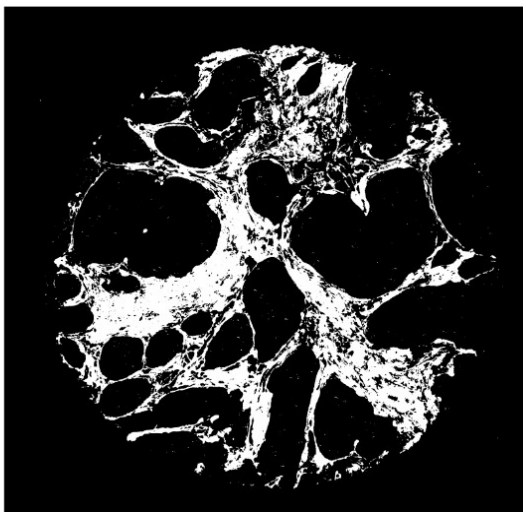
A second convolution operation is performed with a dropout layer added in between. After that, the max pooling operation is applied. It extracts the maximum value of a 2 x 2 region. The result of this group of operations is a downsampled image where the width and the height dimensions are halved and the depth is increased up to the number of kernels used (128 x 128 x 32 in this case). Such a group of operations is repeatedly applied on a new layer, transforming the layer having the shape w x h x d into a deeper layer having the shape w/2 x h/2 x 2d. The deepest layer then contains extracted features that can be used to segment an image. This layer is then upsampled and merged with the previous layer to apply the extracted features. The upsampling process is repeated multiple times to arrive at the layer equal in width and height of the original tile. This layer is then finally deconvoluted to get a segmentation map of shape 256 x 256.

We made a few adjustments to the original U-net architecture to better fit our problem. Firstly, padding was added to each convolutional layer as otherwise the width and height of the deepest layer would become too small. Secondly, we tried to experiment with the architecture's depth (i.e. the number of transformations w x h x d -> w/2 x h/2 x 2d) and found out that the extra layers obtained from increasing the depth level above 4 are redundant. Thus, we settled for a depth 4 after the experiment.

The 43 image-mask pairs were split into 33 training and 10 test samples. The neural network was trained on the training data and validated on the test data. Adam optimizer was used for training with the binary cross-entropy as a loss function. It took about 15 minutes for the model to train on 33 samples. Accuracy (i.e. the proportion of correct foreground / background pixels) metric was used to test the model. All code was written in python using the Keras neural network API [24]. The trained model achieves an accuracy

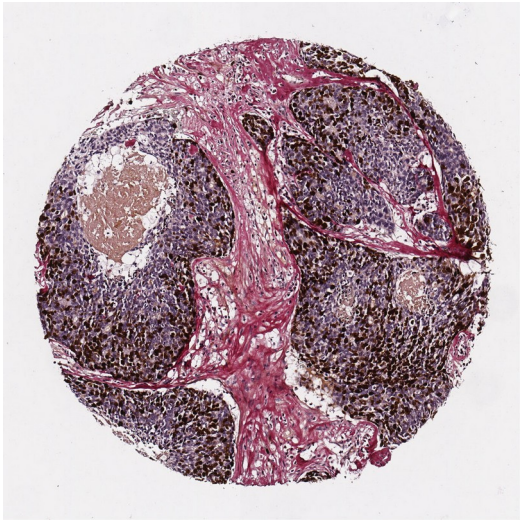
of 92% on the test data. The model was applied on a new larger-scale image for further validation. The larger image was divided into 256 × 256 tiles and the model was applied to each tile to extract a collagen mask. The extracted masks were then merged into a large mask to obtain the result below. The mast correctness was verified by an expert pathologist at the National Pathology Centre, Vilnius.





Binary Classification Problem

After the procedure to extract the collagen from an image is done, the next step is to use it for diagnosis. For this project I was given 93 H&E stained images, which look as follows:



Each image is a biopsy of a breast cancer patient. The size of an image is 1500 x 1500 pixels. Attached to an image is the survival data for that patient:

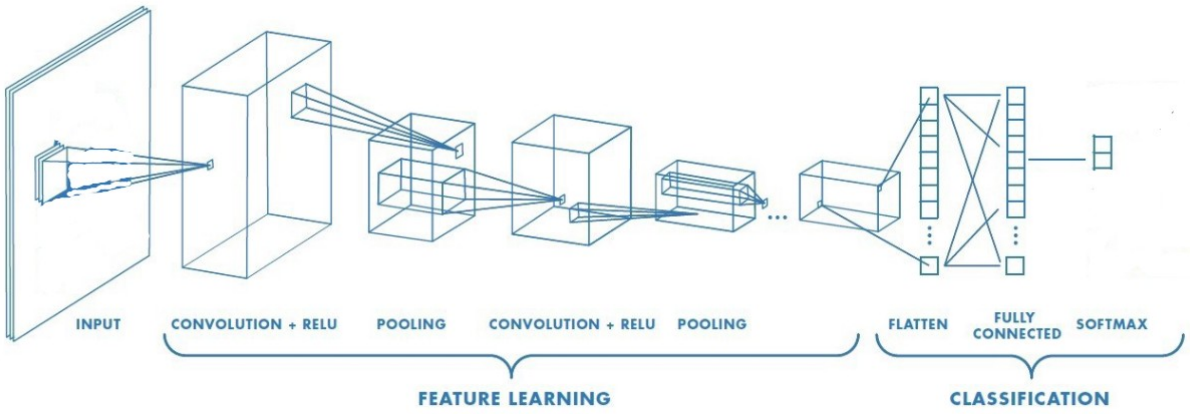
patId	id	ost_m	status01
1	28979	120	0
2	23721	76	1

Survival data contains the number of months a patient was alive from the time the biopsy was taken until the event. The event can be either a patient's death or a censoring event. The censoring event indicates the end of the study. In the example above, the patient nr. 1 was in a study for 120 months after his biopsy was taken. He then left the trial and his current survival status (i.e. dead or alive) is not known. Note that most likely the censored patient is still alive because once per year the doctor calls the patient to get an update on his status. On the other hand, the patient nr. 2 lived for 76 months from the time his biopsy was taken and then died.

The first idea is to assume all censored patients are alive and then classify images into two groups: dead or alive. In this way we turn the data into a binary classification problem: given an image, predict the group to which it belongs. In this data set, there are 21 dead and 72 censored patients. Since the proportion of the sizes between the two groups is imbalanced, we augment the data to make it more balanced. We rotate each image corresponding to a dead patient by 90, 180 and 270 degrees, thus increasing the dead

group size by a factor of 4. This gives us a dataset consisting of 84 dead and 72 (presumably) alive patients.

We are now going to train and test a binary classifier on the augmented dataset. We first use our modified U-net mentioned above to compute a collagen mask for each image. We then build a second convolutional neural network that takes a collagen mask as an input and predicts the binary (dead or alive) output:



The second convolutional neural network applies a convolution operation followed by a max-pooling operation multiple times to learn the features hidden in the image. The final convoluted layer is then flattened into a vector and a dense layer is attached to use the learned features for classification. The dense layer is then joined with the final layer of two nodes, each corresponding to the patient's survival status.

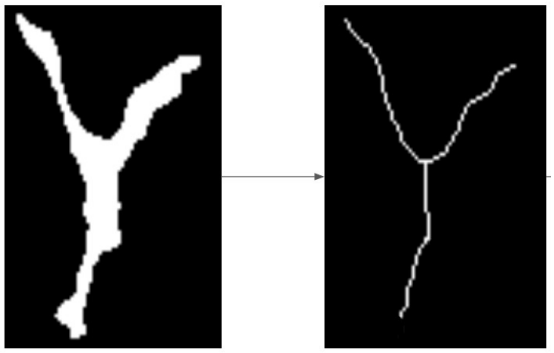
To train the second convolutional network, we split the 156 masks into 77 training samples, 33 validation samples and 46 test samples. We train the model on the 77 training samples and validate its performance on the 33 validation samples. We try a different number of nodes in each layer (e.g. we varied the number of nodes in a final dense layer from 32 to 256) to optimize the validation accuracy. The model that gives the highest accuracy on the validation set is then applied to the test set as a final performance test. We report 55% accuracy on the test set.

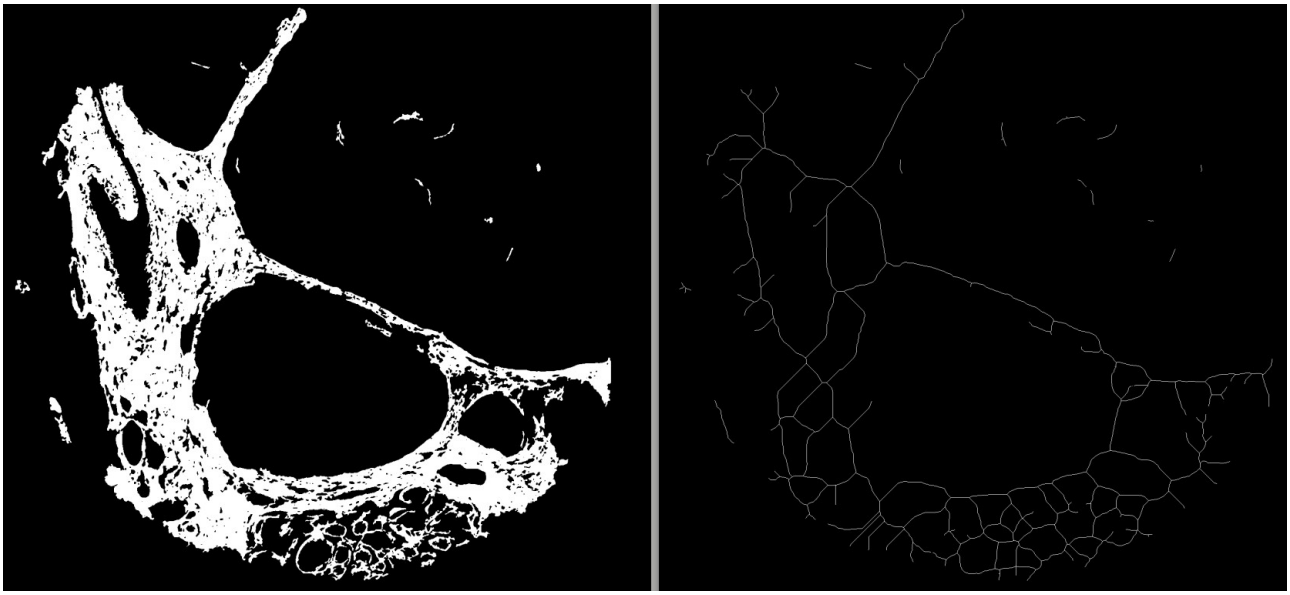
For completeness, we also constructed a convolutional neural network that takes a whole coloured image as an input (without extracting the mask) and predicts the binary survival status as an output. After the model was trained and validated, we report 53% accuracy on the test set.

Special Features Generation

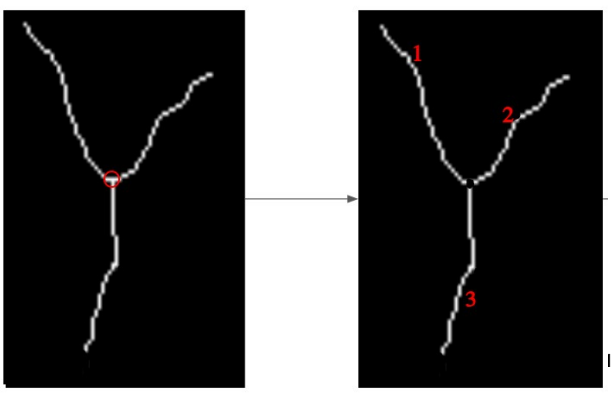
The poor performance of the convolutional neural network used on the whole image suggests that the data is inherently complex enough and the network fails to learn by itself the features present in the image. To address this problem, we decided to compute some hand-crafted features on the collagen mask that could be later used for prediction. For each mask we execute the following pipeline to compute the features (i.e. several numbers):

• A skeletonization function [28] is applied on the mask. The skeletonization function of a shape is a thin version of that shape that is equidistant to its boundaries. This transforms white collagen regions into 1 pixel wide representations and thus the whole mask is transformed into a skeleton tree:

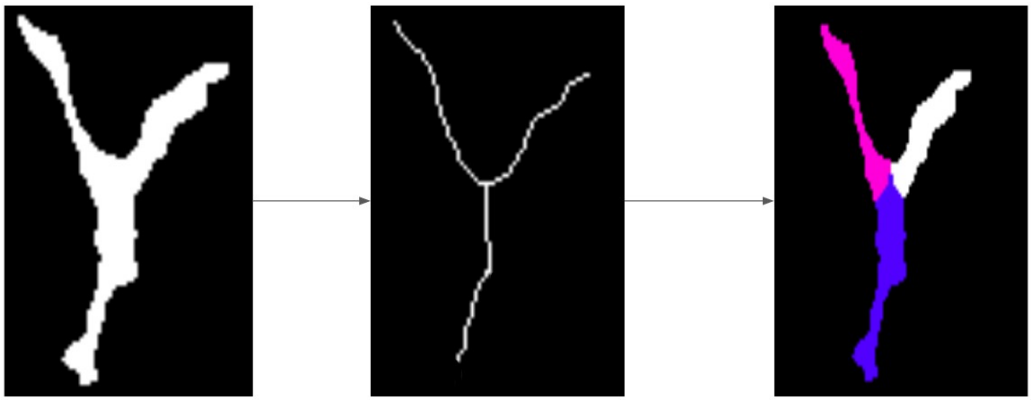




• The whole skeleton tree is chopped into individual edges chopping along the white pixel points that have more than two neighbours. Each edge gets assigned a unique id:



• A breadth-first search algorithm (BFS) [1] is applied to find for each collagen pixel in the mask its closest skeleton edge. Each white pixel in the collagen mask is viewed as a vertex in the graph. Two vertices are connected by an edge of length 1 if the white pixels are adjacent with respect to the 8-connectivity [19]. For each vertex v we search for the closest vertex u that also belongs to one of the skeleton edges. We use the BFS algorithm to find the closest vertex u. We assign the id of the skeleton edge of the vertex u to the vertex v. In this way we divide the whole collagen mask into regions (e.g. pink, white and blue regions are shown in the diagram below):



• For each region we compute 16 statistics:

1) The length. The length of the region is defined as the length of its skeleton edge. To measure the length of the skeleton edge we find one corner in the skeleton (a skeleton pixel that has only one white neighbour). We then traverse the skeleton edge transitioning from the latest visited white pixel to the new unvisited white neighbouring pixel. We add 1 to the total sum if the transition was horizontal or vertical (e.g. if we moved from a pixel having the coordinates (x,y) to the neighbouring pixel having the coordinates (x, y+1)). We add sqrt(2) to the total sum if the transition was diagonal (e.g. from (x,y) to (x-1, y+1)). We return the total sum as the length feature.

2) The width. We define the width as the region area divided by its length. The length is computed as above. The area is total number of white pixels in the region.

3) The curvature. We define it as 1 – D/L. D is the euclidean distance between the two endpoints of a skeleton edge. L is the skeleton edge length measured as in (1). For

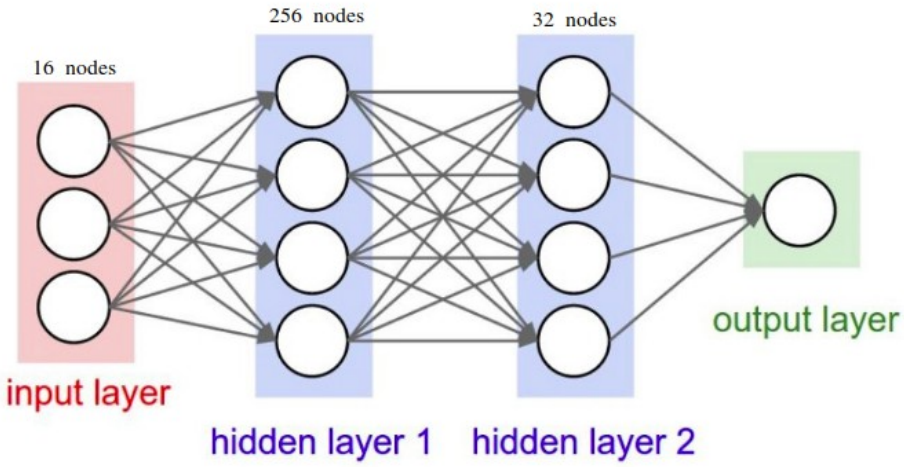
completely straight edges D=L and thus the curvature is 0. Otherwise the curvature is a number in the range (0,1).

4-16) 13 haralick parameters [11,12]. We exclude the 14th due to its instability.

• The final step is to take the average value of each feature. I.e. For each feature, we sum the values over all regions and divide the sum by the number of regions a collagen mask was partitioned into. The result is the 16 numbers (each number corresponding to the averaged feature) computed for each mask.

Binary Classification Using the Special Features

Having computed the features we can now feed them into a neural network. We hope that by eliminating the task for a network to learn the features, it could train to optimize itself for better classification accuracy. The network chosen for this task was a dense neural network with 2 hidden layers:



The input layer has 16 nodes each corresponding to one of the 16 features. The output layer predicts one of the two classes: dead or alive.

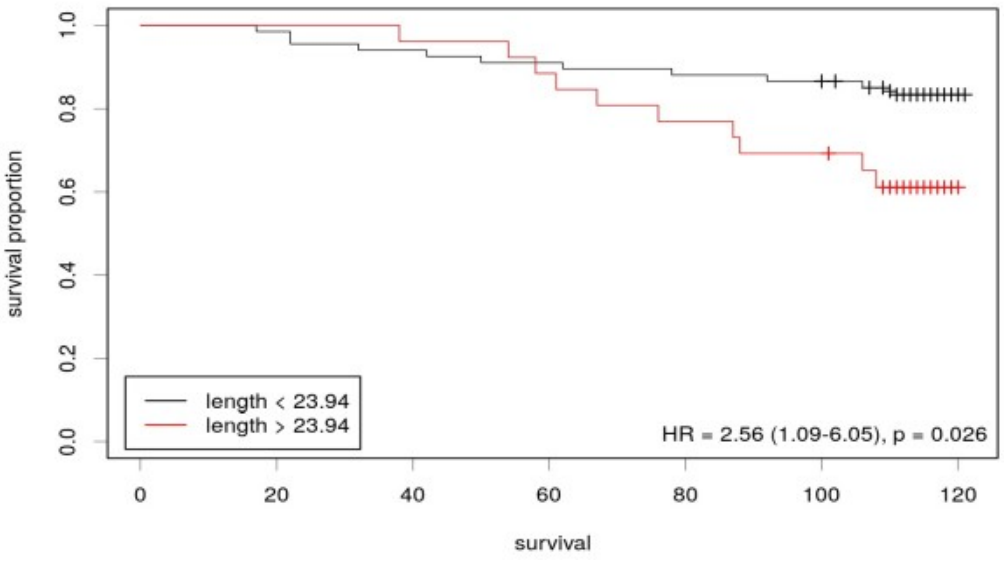
As before, to train the neural network, we split the 156 masks into 77 training samples, 33 validation samples and 46 test samples. We train the model on the 77 training samples and validate its performance on the 33 validation samples. We tried a different number of nodes in each hidden layer (computing its accuracy on the validation set) and settled down for 256 nodes in the first layer and 32 in the second. The final accuracy reported on the test set is 57%.

Regression Problem

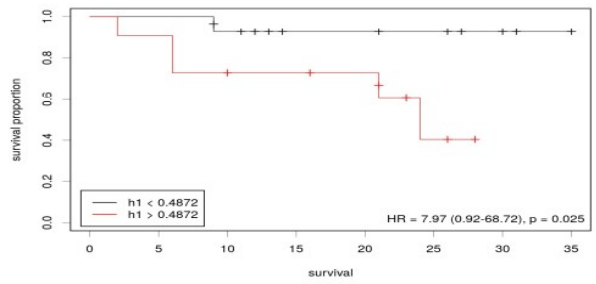
Failing to achieve meaningful results with the classification problem, we now switch to the regression problem [25]. In the regression problem, we want to find a correlation between the images and the continuous survival time (rather than the discrete dead/alive status). Since the continuous survival time is censored, we apply the specific Kaplan-Meier analysis [15] for this purpose:

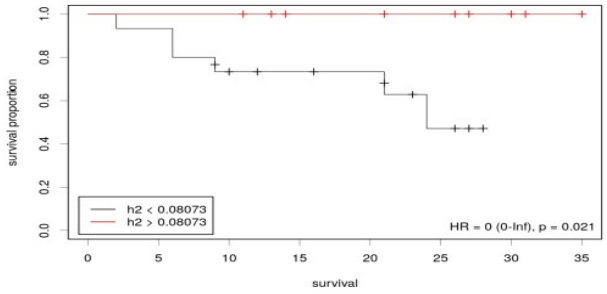
For each of the 16 computed features (i.e. for a list of 16 numbers, one list per feature) we use a Cutoff finder package [2,6] to derive a threshold for that feature. We then partition the patients into two groups based on the threshold. The data of each group can be used to approximate a survival function for that group. The survival function P(t) is the probability that a patient survives for at least time t. Each group has its own true hidden survival function P1(t), and P2(t). We use the log-rank test [17] on these two groups to compute the p-value. The p-value [23] in simple terms tells the probability the same survival function P1(t) = P2(t) could have generated the observed data. Thus, a small p-value indicates that the survival functions P1(t) and P2(t) are unlikely to be the same and thus the 2 groups significantly differ with respect to their survival time.

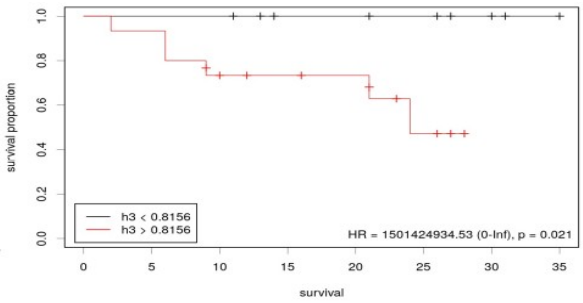
Using 5% as the threshold, we report that only the length feature significantly separates the cohort into 2 groups with the p-value of 2.6%:

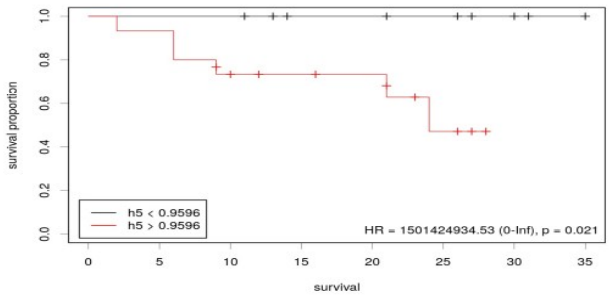


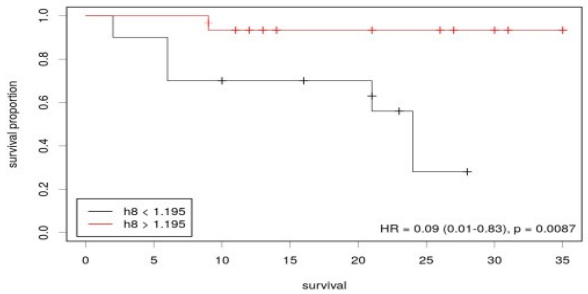
However, although for one specific feature the chance to obtain a significant p-value accidentally is small, the chance for at least one out 16 features to obtain a significant p-value accidentally can be reasonably large. Therefore, to validate the length feature, we apply the same Kaplan-Meier analysis on another dataset. The other cohort comprises 25 large H&E stained images (e.g. 57767 × 45287 pixels), each image per patient. 6 of them are related with the negative outcome (the patient has died), 19 with the positive. We first compute the 16 features mentioned above for each of the patients on the new dataset using the same procedure as before. For each of the 16 computed features we then compute the corresponding p-value. For 10 of these features (h1, h2, h3, h5, h8, h9, h10, h11, h12, h13) the computed p-value scores are significant (less than 5%):

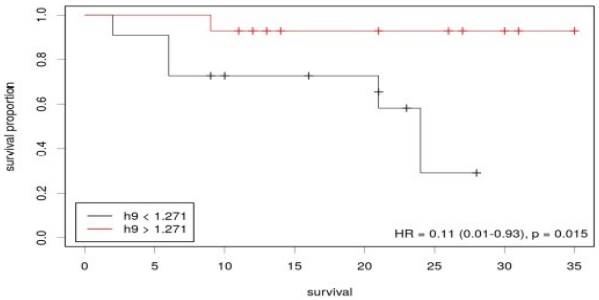


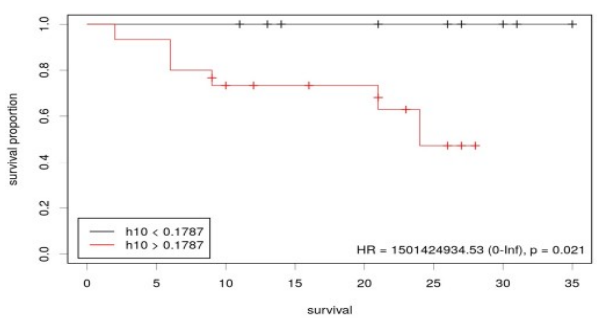


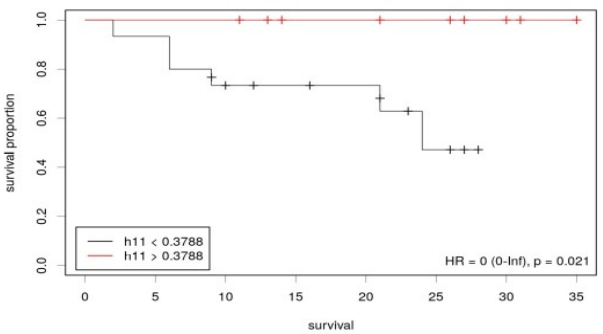


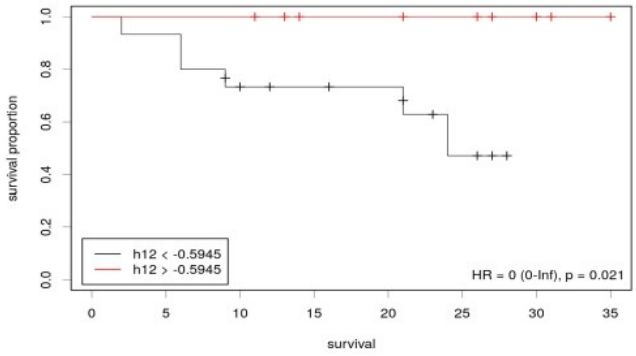


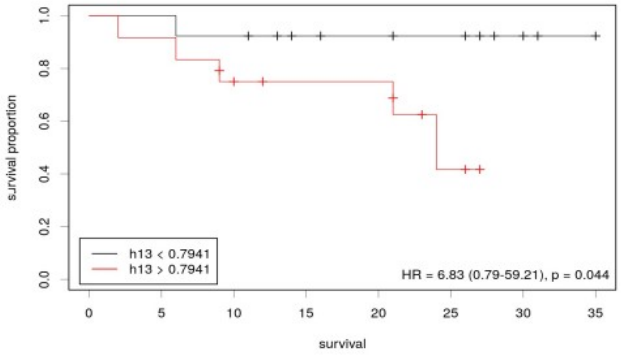










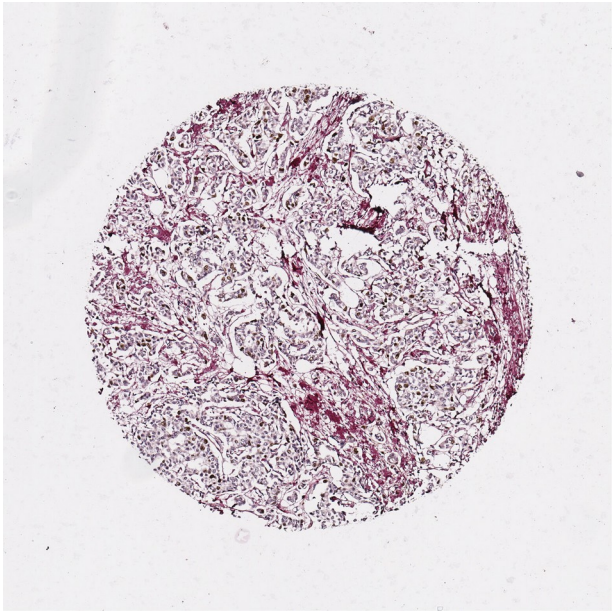


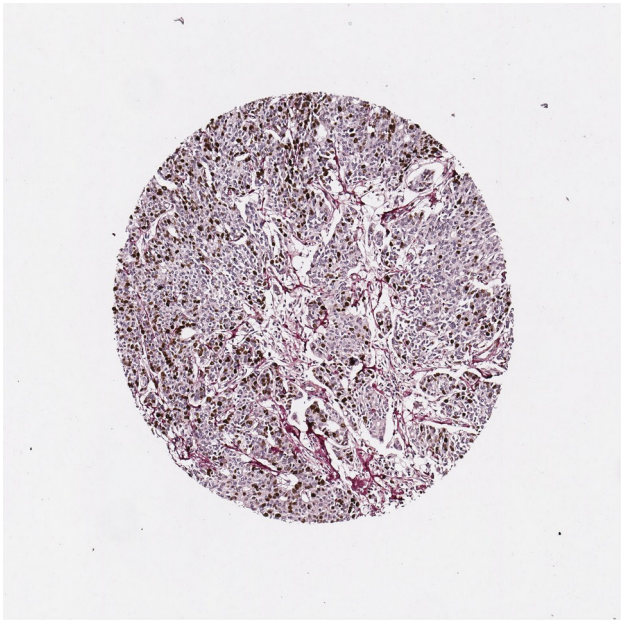
As one can see, only the Haralick parameters (and not the length feature) are significant for this cohort. We conclude that none of our computed features can reliably partition a data set into different survival groups.

DISCUSSION

In this project we wrote some software to analyse H&E stained cancer tissue images. The aim was to present a novel method that could take an image and automatically make a diagnosis for the patient, thus supplementing the TNM classification. The focus was placed on the extracellular protein collagen, indicated in the literature as a promising biomarker.

Our results show that the methods presented in the project do not generate any promising results on the data given. One reason for that might be that the data does not have enough “power” for meaningful results. Below is an example to support this argument:





The patient with the imaged biopsy picture on the left has been alive for more than 114 months after the biopsy was taken. The patient with the imaged biopsy picture on the right has died after 50 months from the time the biopsy was taken. Despite two completely different autcomes, both images look similar.

In comparison, a single image is not enough for a TNM classification. This system requires knowing the size of a tumour (T), the degree of spread to new-by lymph nodes (N), as well as information about metastasis to other organs and tissues besides the lymph nodes (M), none of which are apparent from a single H&E stain image. Also, histological scoring often requires the user to scan around a slide and review many different fields of view.

CONCLUSIONS

Below are the conclusions to the following tasks:

- gather data: a list of patients with information (e.g. survival time) and a digital image per patient;

  The data was succesfully provided by the national state of pathology. As the project demonstrates, this data was not sufficient to make a scientific breakthrough.

- write a method that extracts collagen from the image;

  A convulional neural network can succesfully extract collagen from the RGB H&E stained cancer image.

- compute features using the extracted collagen; use the computed features to make a prediction; evaluate the whole pipeline that makes a prediction;

  The range of features that one can extract from collagen is wide. None of the tested features generate promising predictions, which concludes that the hidden information present in the collagen pattern is too complicated to decipher using the tools of this project.

ACKNOWLEDGEMENTS

Special thanks goes to these people:

Povilas Treigys – project supervisor;
Mindaugas Morkūnas – project supervisor;
Elzbieta Budginaitė – course mate;

REFERENCES

1. Breath-first search algorithm: https://en.wikipedia.org/wiki/Breadth-first-search [September 2019]
2. Budczies Jan, Carsten Denkert. Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization, 2012.
3. Cong L, Yupei Zhao. Tumor size classification of the 8th edition of TNM staging system is superior to that of the 7th edition in predicting the survival outcome of pancreatic cancer patients 2018 (https://www.nature.com/articles/s41598-018-28193-4)
4. Conklin MW, Eickhoff JC, Riching KM, Pehlke CA, Eliceiri KW, Proven-zano PP, et al. Aligned collagen is a prognostic signature for survival in human breast carcinoma. Am J Pathol. 2011;
5. Cox regression: http://www.sthda.com/english/wiki/cox-proportional-hazards-model
6. Cutoff Finder software. http://molpath.charite.de/cutoff/ [December 2019]
7. Drifka CR, Eliceiri KW. Comparison of picrosirius red staining with second harmonic generation imaging for the quantification of clinically relevant collagen fiber features in histopathology samples. J Histochem Cytochem 2016 Sep.
8. Garcia G, Garcia R: A Review on Deep Learning Techniques Applied to Semantic Segmentation
9. Ghosh S, Maulik U: Understanding Deep Learning Techniques for Image Segmentation [July 2019]
10. H&E stain. https://en.wikipedia.org/wiki/H%26E stain [September 2019]
11. Haralick features definition. http://murphylab.web.cmu.edu/publications/boland/boland_node26.html [September 2019]
12. Haralick features in python. https://mahotas.readthedocs.io/en/latest/features.html [September 2019]
13. Image Segmentation: https://en.wikipedia.org/wiki/Image_segmentation [March 2020]
14. Jeremy S. Bredfeldt, Kevin W. Eliceiri1. Automated quantification of aligned collagen for human breast carcinoma prognosis. J Pathol Inform, 2014;
15. Kaplan-Meier estimator. https://en.wikipedia.org/wiki/Kaplan–Meier estimator [February 2010]
16. Krizhevsky A, Hinton G: ImageNet Classification with Deep Convolutional Neural Networks
17. Log rank test: https://en.wikipedia.org/wiki/Logrank_test [April 2020]
18. Pavithra V, Nambiar S. Tumor-associated Collagen Signatures: An Insight, 2017 (https://pdfs.semanticscholar.org/505d/bbe18909bd8b8806e9cf4a4655d997ec33a3.pdf)
19. Pixel connectivity types. https://en.wikipedia.org/wiki/Pixel_connectivity [September 2019]

20. 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.
21. Provenzano PP, Inman DR, Eliceiri KW, Trier SM, Keely PJ. Contact guidance mediated three-dimensional cell migration is regulated by Rho/ROCK-dependent matrix reorganization. Biophys J. 2008.
22. 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.
23. P-value: https://en.wikipedia.org/wiki/P-value [April 2020]
24. Python Keras API. https://keras.io/ [September 2019]
25. Regression analysis: https://en.wikipedia.org/wiki/Regression_analysis [April 2020]
26. Ronneberger O, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.
27. Rushing C, Bulusu A, Pang H. A leave-one-out cross validation SAS macro for the identification of markers associated with survival. Comput Biol Med, 2015
28. Skeletonization function. https://scikit-image.org/docs/dev/auto_examples/edges/plot_skeleton.html [September 2019]
29. Strupler M, Schanne-Klein MC. Second harmonic imaging and scoring of collagen in fibrotic tissues. Opt Express 2007 Apr;15(7):4054-4065.
30. Tissue microarray. https://en.wikipedia.org/wiki/Tissue_microarray [September 2019]
31. TNM staging system [April 2020]

SUMMARY

Marius Latinis

Systems Biology Master's Programme, Vilnius University

Collagen-based Survival Analysis in the Hematoxylin and Eosin Stained Histological Images

Visual examination of a tissue is a major part of a cancer diagnosis. A trained pathologist inspects a tissue's sample that is often stained using the Hematoxylin and Eosin to make a prediction. However, manual image inspection is prone to error. This project aims to propose a new method that can inspect the H&E stained images automatically and generate a diagnosis for the patient. The method is based on the protein collagen, that is visualised as a pink colour on the image. First, the collagen mask is extracted from the image using the convolutional neural network. Next, several features are computed from the mask. The features are then fed into another network that trains itself using the given survival data to make a prediction. The prediction results are reported. We conclude that this methodology brings no significant improvement compared to making the prediction randomly.

SUMMARY IN LITHUANIAN

Marius Latinis
Sistemų Biologijos Magistras, Vilniaus Universitetas

Kolegeno struktūrų analizė išgyvenanmumui hematoksilinu-eozinu dažytuose histologiniuose vaizduose

Regimoji vėžinio audinio analizė yra vienas iš pagrindinių vėžio diagnostikos elementų. Patologas ištiria audinio mėginį, dažniausiai nudažytą Hematoksilinu ir Eozinu, kad atliktų diagnozę. Deja, rankiniame vaizdo tyrime gali įsivelti klaidų. Šio projekto tikslas yra pristatyti naują metodą kuris galėtų automatiškai (t.y. kompiuterio pagalba) ištirti H&E nudažytus vaizdus ir nustatyti paciento diagnozę. Metodas paremtas baltymu kolagenu, kuris H&E nudažytose nuotraukose atvaizduojamas rusva spalva. Visų pirma, kolageno forma yra ištraukiama iš spalvotos nuotraukos. Tada, kolageno forma yra panaudojama išskaičiuoti keletą požymių. Tie požymiai yra įvedami į kitą neuroninį tinklą generuojantį diagnozės spėjimą. Spėjimų rezultatai yra raportuoti šiame projekte. Padaryta išvada, kad projekte naudota metodika neatneša geresnių rezultatų nei darant atsitiktinį spėjimą.