UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Assessment of histoQC as a quality control tool for whole slide images of brain tumour biopsies

First examiner:
Artem Kaznatcheev

Second examiner:
Sanne Abeln

Candidate:
Nanda Pellikaan

In cooperation with:
Sebastian Waszak
Birgit Kriener
NCMM Oslo

July 12, 2023

Abstract

Cancer diagnosis and treatment is currently performed using histopathology, which is time consuming and labour intensive. This process can be automated using digital pathology, if the quality of the input slides is good enough.

Slide quality is important, due to the fact of differences in staining colour, that originate from the variety of staining processes performed at the different hospitals. Furthermore, in the process of staining slides, artifacts can occur.

So, when bad quality slides are used as input for model building, biased models will result. In this paper I answer the research question 'Is histoQC a good tool for assessing the quality of the whole slide images in the cancer genome atlas glioma dataset?'.

The results show that histoQC in its default configuration is not accurate in detecting artifacts on slides. However I found that the output metrics histoQC generates are a good input for similarity analysis.

Knowing that histoQC is a good tool to find similarities in the input slides, further research can be done to find out if the similarity analysis can predict the robustness of a model trained on a specific set of the data.

Contents

1. Introduction

Currently, cancer diagnosis and grading is performed by a trained pathologist. In this process tissue samples are obtained during a biopsy or surgical resection, after which preparation of the slide is performed through staining the slide with haematoxylin and eosin (H&E). The slide is then examined by a human expert using an optical microscope. This visual examination is a repetitive process which can be automated, in order to reduce costs and turnaround time. [1], [2]

Digital pathology is the digitisation of the traditional diagnostic process of analysing cells and tissue with a microscope via whole slide image (WSI) scanners, computer screens and data science. [3] Using digital pathology, the pathologist digitises the H&E stained glass slide using a scanner. In this digitisation process, the same magnification is used as when analysing the slide with a microscope. After scanning the slide, the pathologist can either analyse the slide manually on the computer screen or (parts of) the analysis can be automatised by the computer. [4]

There are data quality challenges in automating the histology workflow. First, the process of staining and digitizing an H&E slide happens differently in every hospital. This results in darker or lighter images. Second, in processing the slides the pathologist sometimes writes a note on the slide or draws a circle around an area of interest. Third, while processing bubbles or dust can appear on the slide. Last, not every hospital has the same scanner. [5] These problems can result in biased models that can predict the hospital at which the slide is generated really well. [6] Since some hospitals process more slides containing one kind of cancer and others more of another kind, this can result in bad predictions about the cancer types presented on the slide.

Andres Janowczyk et al. developed the histoQC module to assess the quality of whole slide images (WSI) and reduce the bias in models trained on the

slides. [5], [7] HistoQC assesses the slide by using various build-in modules that use statistics, classification models, convolution operations and comparison of image values to the average values in the image. In this process histoQC outputs metrics about the slides, figures containing the information about the artifacts in the slides and masks that can overlay the original slide to subtract tissue on the slide that can be used for further analysis. All these outputs can be used to check the quality of the WSI and select similar slides for further analysis, to reduce bias. This module has already generated good results for assessing the data of H&E stained images of kidney biopsies [5]. Furthermore, Janowczyk has done research to compare the results of histoQC with assessing the slides by histopathologists and found that they agree 95% of the time [7]. These two examples indicate histoQC is a good tool for assessing the quality of WSI and detecting artifacts on them.

In this paper, I will answer the research question 'Is histoQC a good tool for assessing the quality of the WSIs in the cancer genome atlas (TCGA) glioma dataset?' and the subquestions 'Can histoQC accurately detect artifacts on the WSI in the TCGA glioma dataset?' and 'Can the output metrics of histoQC be used for performing a similarity analysis on the WSI in the TCGA glioma dataset?'. In order to find the answers to this questions I will first describe the background on brain tumours, histopathology and slide quality. Second, I will describe the data. Third, I will propose the methods for assessing the quality using the histoQC package. Last, I will discuss the results and provide options for further research.

2. Background

2.1 Brain Tumours

In this paper I will focus on primary brain tumours, which are tumours in the brain and central nervous system (CNS). [8], [9] There are different types of primary brain tumours known, from which I will focus on the malignant brain tumours known as gliomas. [8] A glioma is a type of tumour that originates in the glial cells, the non-neural cells that do not produce electrical impulses.[10]–[12] Around 75% of the malignant primary brain tumours are gliomas. [8] Among gliomas, the following types are known and named based on the type of cell with which they share histological features.

- *Astrocytoma's* originate from star-shaped glial cells that are located in the cerebrum. An astrocytoma usually does not spread outside the CNS and therefore will not affect other organs [13].They have defined borders and develop slowly. They are known as grade I or grade II tumours and are the least aggressive type of brain tumours.
- *Oligodendrogliomas*, are tumours that originate from the oligodendrocytes of the brain or from glial precursor cells and is mostly found in the frontal lobe [14]. Oligodendrogliomas are more aggressive tumours and are considered grade III tumours.
- *Glioblastomas* are the most aggressive type of brain cancer and are classified as grade IV brain tumours. The cellular origin of this type is not known. However, recent research show that astrocytes, oligodendrocyte progenitor cells and neural stem cells could all serve as cell origin. [15]–[19]

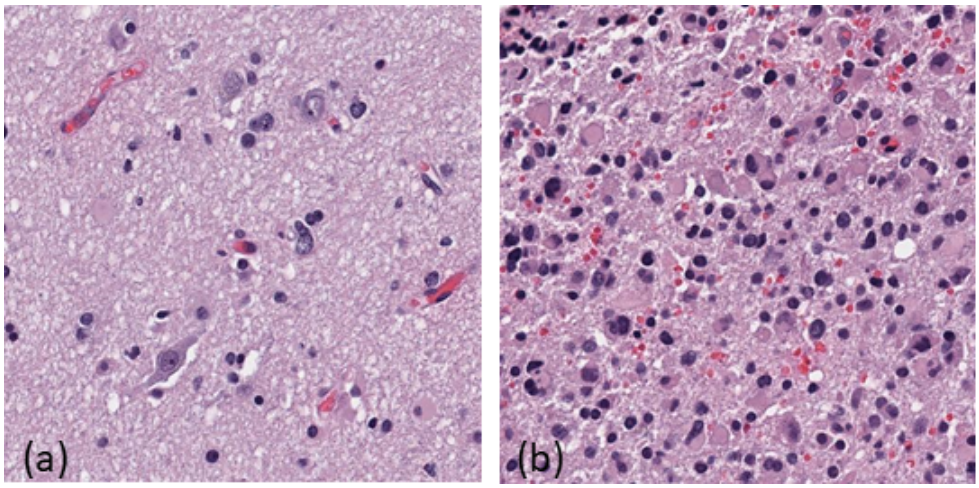

Figure 2.1: Histopathology Image: from (a) healthy brain tissue and (b) malignant brain tissue. Both images represent a 256x256 µm representation of a WSI. I created these images from WSIs in the TCGA dataset.

2.2 Histopathology

Histopathology is the field of human tissue analysis for a specific disease and the histopathology images as described here are currently the principle information source for cancer diagnosis and prognosis. These histopathology images or whole slide images (WSI) are prepared in a process in which the tissue sample is stained on a glass slide using haematoxylin and eosin (H&E). H&E staining results in dark purple coloured nuclei, with the other tissues coloured pink. Under the microscope the tissue is analysed to diagnose or grade brain tumours. [20]

An example of a slice of a WSI is shown in figure 2.1, on the left side an example of healthy brain tissue is visible, where on the right side malignant tissue is present. Comparing these two parts of a WSI, the differences are clear. In the healthy tissue, less cell nuclei (purple dots) are present on the slide, than in the malignant tissue. Furthermore, in the healthy tissue the nuclei are more round and evenly distributed over the slide. The variation that is visible in the healthy tissue comes from the different cell types present in the brain tissue. In the malignant tissue, more cell nuclei are present, with lots of variation in their shape and size. This variation can be both explained by the different cell types present in brain tissue as by the abnormal cell growth of tumours.

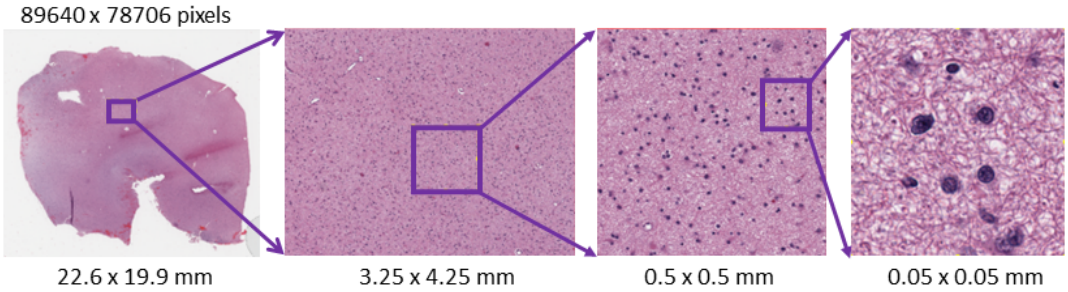

Figure 2.2: A WSI, the whole slide (a) is 22.5x20mm in steps the WSI is zoomed in untill almost pixel level in the lower left corner (d) which has a size of 0.05x0.05mm, and in which the cell nuclei and other cell tissues are finally separable.

In histopathology, for every biopsy taken, the pathologist needs to analyse the WSI, which are at least 15 x 20 mm in size, where the representations in figure 2.1 are only 256 x 256 µm. So you would need 59 by 78 images like figure 2.1 to produce a typical WSI. In figure 2.2 an example of a typical WSI is shown. As this image shows, a significant zoom-in is necessary to distinguish the cell nuclei and other cell tissues. Combining the big size of the data with the possibility of both healthy and malignant tissue being present, makes histopathology a highly skilled and time-consuming process that is prone to human error.

2.3 Slide Quality

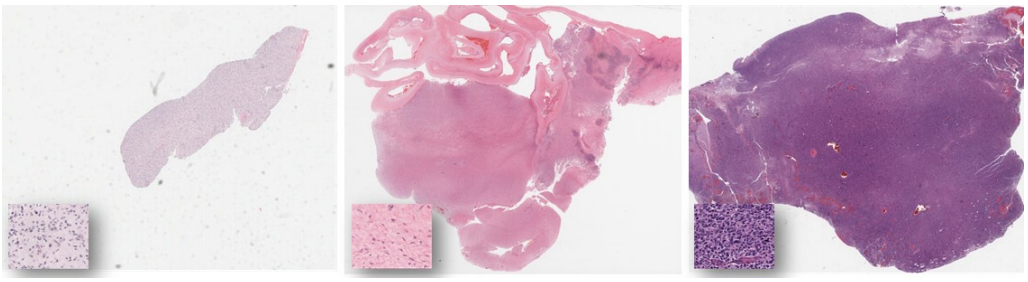

Figure 2.3: The variation in colour in WSIs stained with H&E. The colour variance can most likely be explained by the differences in staining process of the hospitals that contributed their data to the TCGA dataset. As shown in the zoom-in at the lower left corners, the colour differences are not explained by the nuclei density. I created these representations from the WSIs in the dataset.

In order to do successful analysis on WSIs, the quality of the slides is

important. With high quality slides, prepared via the same staining process, the data from multiple hospitals will be comparable. However, hospitals currently have their own staining process, which can lead to variations in colour intensity between the slides from different hospitals, as shown in figure 2.3. The zoom-in representations in the lower left corners show that the colour variation can not solely be explained by the nuclei density in the WSI, since the other cell materials also differ in colours between this three examples.

Furthermore, sometimes artifacts are present on the slides. Artifacts can originate from the glass slide, the tissue and from the scanning of the slides. These artifacts influence the computational pathology process by introducing information that does not contain tissue and therefore can reduce the accuracy of models trained on the slide. [5] In figure 2.4 examples of artifacts are shown. Figure 2.4 a, b, c and d show the artifacts caused by the glass slide. For example in (a) a slide that is dirty. The dirt can be recognised by the grey parts on the slide. In (b) a slide is marked with pen. Sometimes the pathologist highlights a part of the slide with a marker or writes something on the slide. In (c) a cover slip is visible on the left side of the slide as a black vertical stripe. This artifact comes from the top glass (cover) of the glass slide slipping. In (d) we see an air bubble. This comes from air that is caught between the bottom and top glass of the slide.

In figure 2.4 e an example of a tissue section artifact is shown. Tissue section artifacts are caused by the tissue on the glass slide. In case of (e) in the middle of the slide an intense, dark purple area is visible, caused by the tissue being folded on the slide. In figure 2.4 f and g we see artifacts that originate from the scanning process. In (f) the slide contains a blurry section caused by the scanner being out of focus when capturing that part of the slide. In (g) grid noise is present, which is visible by the vertical lines on the slide. This artifact is caused by the way the scanner moves over the slide.

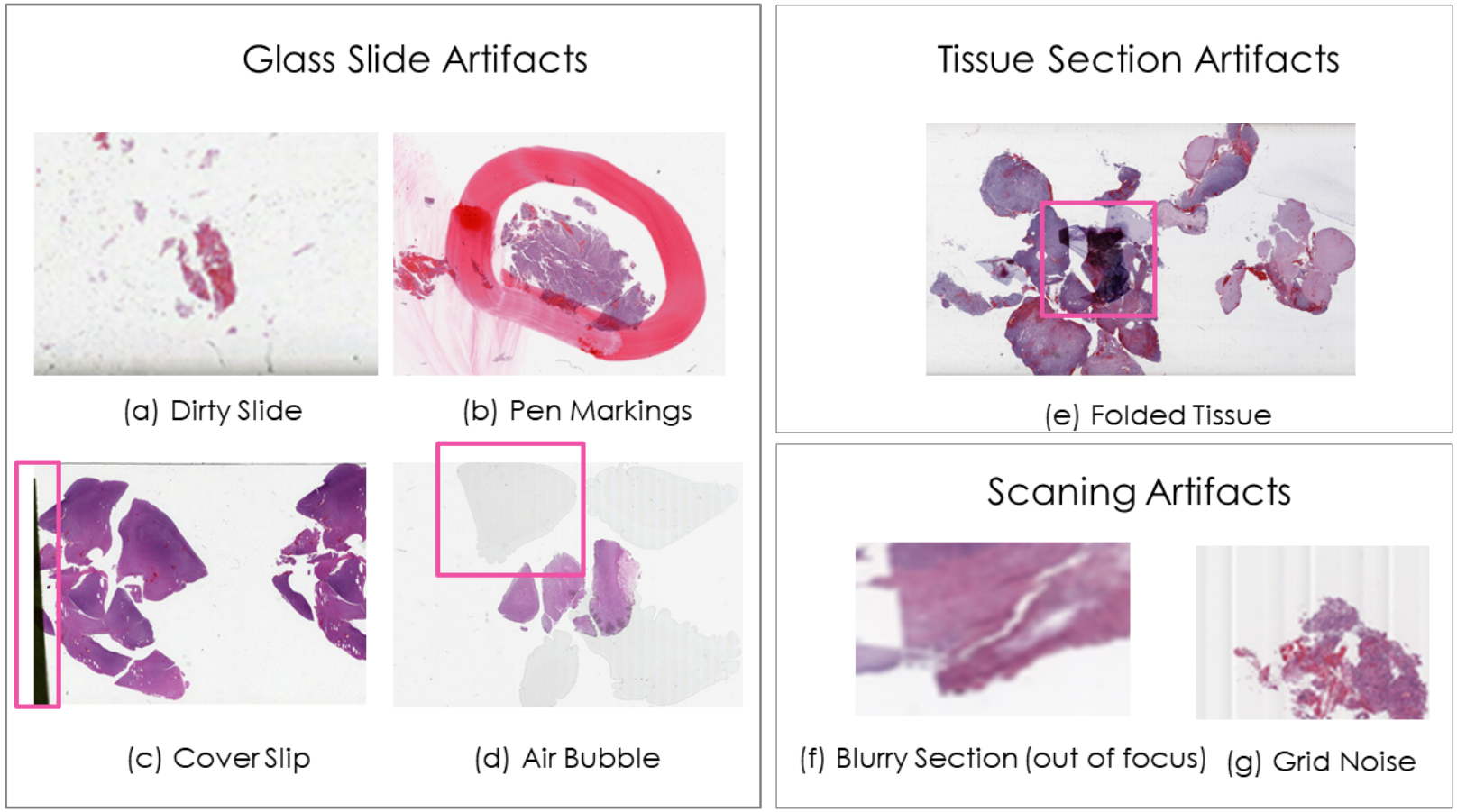

Figure 2.4: Common artifacts on WSI. (a) dirty slide, on the slide dirt is presented next to the tissue. (b) pen markings, the pathologist wrote something on the slide or circled the region of interest. (c) cover slip, the top glass of the slide is visible. (d) air bubble, there is air between the bottom and top glass of the slide. (e) folded tissue, by preparing the slide the tissue got folded represented as an intense, dark coloured tissue part. (f) blurry section, while scanning the slide the scanner got out of focuss. (g) grid noise, scanning lines are represented in this example as vertical lines in this slides. All these examples are gathered from the TCGA glioma dataset.

3. Data

The data for this project was retrieved from the Cancer Genome Atlas (TCGA). It consisted of whole slide images (WSIs) of brain tumours, containing 1704 WSIs of low grade gliomas (LGG) and glioblastoma multiforma (GBM). [21]

The data is retrieved from multiple hospitals, more specifically 41 hospitals for the brain tumour data, and the histopathology process can differ a little. Meaning that the way the tissue is stained on the slide, can differ between the hospitals. This will result in a variation in the intensity of the colours of the WSI, as shown in figure 2.3. Another reason for colour space artifacts can be caused by the use of different scanners in the different hospitals. Furthermore, the data is not obtained for research goals, but for clinical practice resulting in the possibility of artifacts on the slides, as shown in figure 2.4.

4. Method

The aim of this paper is to assess if histoQC is a good tool for quality control of whole slide images (WSIs) from brain tumour tissue in the Cancer Genome Atlas (TCGA). Since previous research indicates that histoQC is a good tool for quality control in H&E stained WSIs of kidney biopsies [5] and that the results of histoQC are in line with assessing WSIs by a histopathologist [7] I choose to use histoQC for the quality assessment of the TCGA glioma dataset.

HistoQC is a quality control tool that is able to analyse WSIs stained by various processess, but is tested most on H&E stained WSIs. In the analysis of the WSIs histoQC first trains a classification model on templates to detect pen markings and cover slips on the WSIs. After that it starts analysing the WSI using statistical measures, convolution operations and the classification model it trained. For a complete list of the modules that are built-in in histoQC see appendix D. It is possible to change histoQCs configuration file to meet the requirements of the project. In this project I used the default configuration file to do the analysis, which means that the classification models are trained on templates provided by histoQC. [22]

Since I divided the aim of the paper in two sub-questions, I will describe the methods used to answer the first sub-question 'Can histoQC accurately detect artifacts on the WSI in the TCGA glioma dataset?' in the section *4.1 artifact detection* and the methods used to answer the second sub-question 'Can the output metrics of histoQC be used for performing a similarity analysis on the WSI in the TCGA glioma dataset?' in the section *4.2 similarity analysis*.

An overview of the complete workflow can be found in figure 4.1

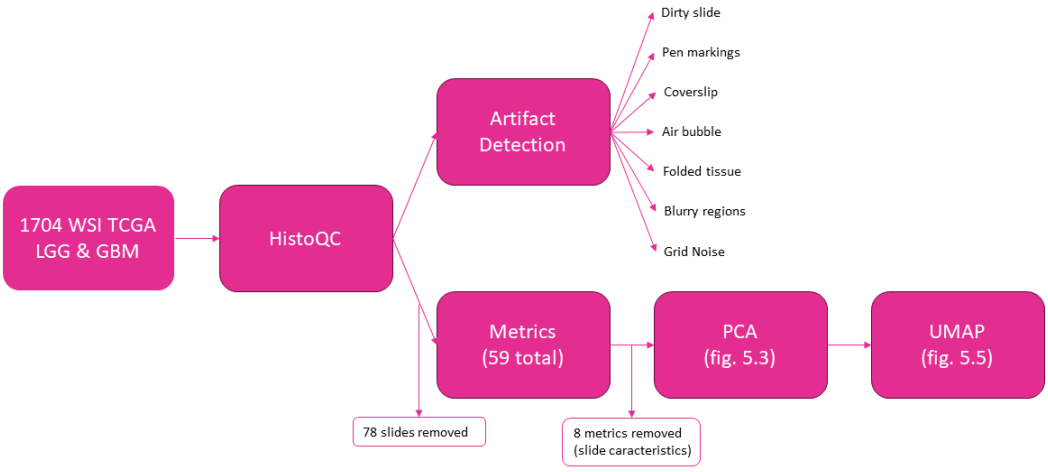

Figure 4.1: Workflow

4.1 Artifact Detection

In the first part of the project, the histoQC module is used to detect artifacts on the WSI of the TGCA dataset. The main goal is to detect the artifacts as described in figure 2.4. In order to detect the artifacts in the WSIs, it is necessary to include the right modules in the configuration that is used for the analysis. In this case I want to detect dirt, pen markings, cover slip, air bubble, folded tissue, blurry sections and grid noise. The default configuration I will use for this project is able to detect all these artifacts using the methods described in table 4.1. For some of the artifacts specific modules are available, namely for pen markings, coverslip detection and blurry regions, the other artifacts are detected based on their differences in the contrast measures compared to the rest of the slide. After running the histoQC analysis, masks are created with the parts of the slide containing tissue that can be used for overlaying with the original slide, to remove the artifacts.
To check the accuracy of histoQC in detecting the artifacts, I manually examine all 1704 WSIs to check if there are pen markings, cover slips and large blurry areas on the slides.

.

Artifact	Module	Description
Dirty Slide	getIntensityThresholdPercent	The getIntensityThresholdPercent operation compares the pixel values to a threshold. If the pixel value is above the threshold, the pixel is detected as dirty.
Pen mark	ClassificationModule	Before analysing the slides a random forest classifier is trained on an example image and a mask. During the analysis histoQC checks if the image has a comparable mask and therefore needs to be classified as containing a pen mark.
Cover slip	ClassificationModule	Like for the pen marking for the coverslip an example slide and mask is provided on which a random forest classifier is trained. During analysis the slide is classified as having a coverslip or not.
Air bubble	getIntensityThresholdPercent	The getIntensityThresholdPercent operation checks the image pixel-wise and compares it to a predefined threshold. If the value is above the lower threshold and below the upper threshold the area is marked as containing an air bubble.
Folded tissue	getIntensityThresholdPercent	The getIntensityThresholdPercent checks the slide pixel-wise to compare the values to the predefined threshold. If the value of the pixel is higher then the set threshold, the tissue at that pixel is marked as folded.
Blurry section	identifyBlurryRegions	Searches the images for regions with a certain threshold for blurriness with a predefined threshold. If the pixel is recognised as blurry and has a radius bigger than the set radius, it is marked as blurry.
Grid noise	getIntensityThresholdPercent	The getIntensityThresholdPercent checks the slide pixel-wise and compares it to the predefined thresholds. If the value is higher than the lower limit and lower than the upper limit it is recognised as grid noise.

Table 4.1: Artifacts that are detected, including the module used and a description how the module works.

4.2 Similarity Analysis

Next to artifact detection and creating the masks of useful tissue, the histoQC module also returns metrics of the WSIs. These metrics contain information about the basic information/statistics about the slides, like the scanner used to create the slide, the magnification at which the slide is digitized, the amount of pixels in the WSI, the height and width of the WSI and the mpp (magnification per pixel) in x and y direction. Furthermore, the metrics contain information about the slides colours and artifacts. A full list of the metrics histoQC produced using the default configuration can be found in appendix C, in this table the first 10 rows represent the information about the slide and the other 51 rows contain information about the slides' colours and artifacts.

These metrics about the colours and artifacts in the slides, which are 51 metrics, are used to assess the similarity of the WSIs in the dataset. These metrics contain among others information about the percentage of the slides covered with artifacts, the brightness and darkness of the slides, the amount of contrast and comparisons of the RGB distribution in the slides to a template.
The output table of histoQC containing the information regarding these variables is used as input for the similarity analysis. The similarity analysis is performed in python using a principal component analysis (PCA). The principal components that are created in the analysis give an indication about how much variance in the slides is explained by them. Slides with similar principal component values are likely more alike based on the metrics that are used as input. Based on the results of PCA, the structure of the data is studied and the similarity between the slides. Slides that are close to each other in the PCA results are likely to be similar, where slides far away are likely to be different.
After the PCA in which only 2 of the 51 dimensions in total are analysed, the PCA results will be used as input for an UMAP analysis. UMAP is a similarity analysis method in which multiple dimensions can be reduced to two dimensions. By using UMAP therefore all the PCA results can be captured in only two dimension, meaning that the output plot will better represent the similarity between the slides. The similarity or differences between slides in UMAP analysis can be described based on the distance, slides further apart from each other in the plot are less similar than slides close to each other.

5. Results

For the analysis the default configuration file is used, resulting in 59 metrics as output. From the original 1704 WSI that were used as input data, histoQC was not able to detect tissue suitable for analysis on 4.6% of them due to artifacts or too little contrast between the tissue and background. This means that 78 slides had no tissue suitable for analysis on them and therefore are excluded from further analysis.
In a subsample of 35 of these slides on which histoQC did not detect suitable tissue, I manually checked if indeed the tile I have from a piece of the slide had no high quality tissue on it. In 14% of the checked images, the tile was made out of tissue that could be used for analysis. This means that from the 78 slides which histoQC found not suitable for further analysis 11 slides most likely will be suitable for further analysis.
On average histoQC detected 0.05% of the pixels on the whole slide to have tissue suitable for analysis, with a range of percentages of the pixels in the WSIs that are suitable for analysis is between 0% and 0.27%. The average number of pixels used for analysis by histoQC is 1.6 million, where an average WSI is 80,000 x 60,000 pixels in size meaning that it consists of 4.8 billion pixels. This numbers added together indicate that the numbers are reasonable, but nevertheless it is a really small percentage of pixels, even when the large amount of background pixels is taken into account.

5.1 Artifact Detection

In histoQC there are metrics for the pen markings, coverslip edge and blurry regions, which can be used to calculate the percentages of slides containing this artifacts. For the complete TCGA glioma dataset of 1704 slides, histoQC detected a pen marking on 3.4% of them, a coverslip on 95.4% and a

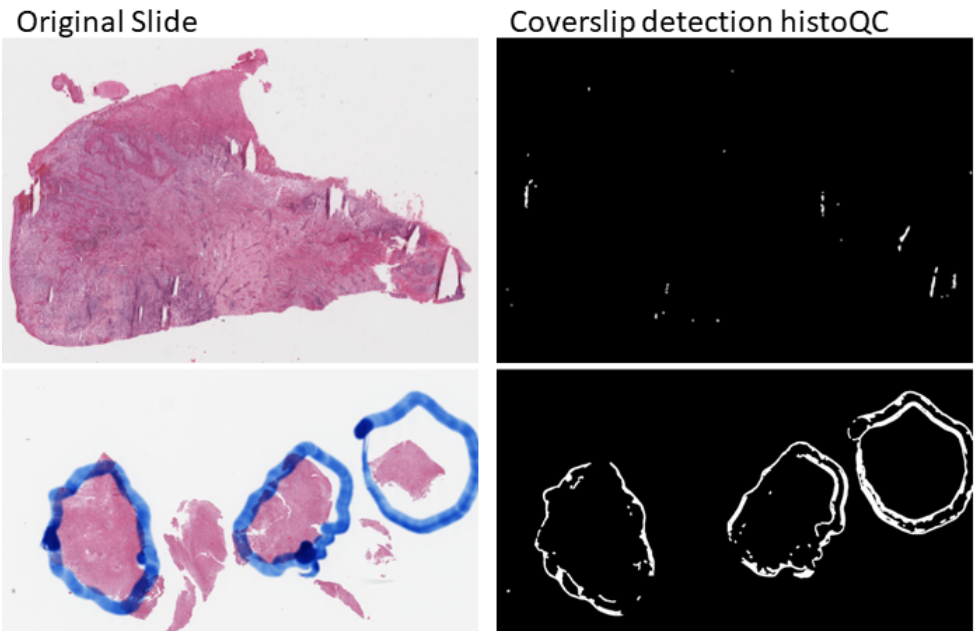

Figure 5.1: Example slides at which histoQC detects a coverslip, but there is none visible in the original slide. In the upper image the borders of the tissue are detected as coverslip and in the lower image the borders of the penmarking are detected as a coverslip.

blurry area bigger than the mean blurry area in 18.5%. The number of slides containing a coverslip is high, due to the fact that histoQC also detects coverslips on slides that don't have them, as shown in figure 5.1. During the analysis of the slides with a blurry region, I found that histoQC detected a blurry region in 99.6% of them. This is because almost all of the slides contain a little blurry region on them at the edge of the tissue, caused by the fact that the tissue density is lower making it harder to keep the scanner in focus. Therefore the number of slides with a blurry region that is bigger than the average area of the blurry region on the slides are measured, resulting in 18.5% of the slides containing a large blurry region.

I manually checked the results for pen markings, cover slips and blurry areas and found that 6% of the slides contains pen markings, 7.7% of the slides contains a cover slip and 8.4% of the slides contain a blurry area that is visible with bare eye. The results of histoQC and the manual check are shown in table 5.1. With the numbers next to each other, the differences in number for the coverslip are really big, which can be explained as described above and in figure 5.1.

Artifact	HistoQC	Eyeballing
Pen mark	3.4%	6.0%
Cover slip	95.4%	7.7%
Blurry	18.5 %	8.4%

Table 5.1: Results of the histoQC artifact detection and the manual check of all 1704 WSI. The difference for the cover slip is really big and can be explained by histoQC detecting cover slips at tissue borders and other artifacts.

In figure 5.2 an example of the visual output of the histoQC results of the artifact detection for the slides represented in the background section, figure 2.4, is shown. In the left column the original slides are represented, in the middle the artifact detection (the grey/black area), the tissue suitable for analysis (the pink area) and the background (in green) are found and on the right the mask that is created and can be used to overlay the original slide in order to extract the tissue suitable for analysis.
When reviewing the slides and the artifact detection visible in figure 5.2, we see in (a) that next to the dirt, also parts of the tissue are masked out resulting in only a small section of tissue suitable for analysis. In (b) the pen marking is detected and also some other artifacts at the borders of the tissue. In (c) the cover slip is detected and masked out, but the tissue on the slide is all selected as suitable for analysis. In (d) the air bubbles are detected, as shown by the colour difference in the green background in the middle column, however this has no effect on the tissue that is suitable for analysis. In (e) the folded tissue near the middle of the slide is detected and also some other parts of the tissue are masked out. In (f) we can see in the middle and right images that none of the tissue in the slides is selected as viable tissue for analysis. At last in (g) we can see that the horizontal lines or grid lines in the original slide are detected by histoQC, but do not effect the tissue that is suitable for analysis. In almost all of these examples, more than one artifact is present, resulting in parts of the tissue masked out. Some of the artifacts are clear from the original slide, but others are hard to detect with bare eye. From this select sample of slides, it shows that histoQC is capable of detecting artifacts and creating a mask for subtracting the tissue that is suitable

for analysis. However, there are no distinct modules in histoQC that create metrics on the grid noise, air bubble, folded tissue or dirt. So, it might be an accident that these artifacts are removed from the slide in these examples. Furthermore, we found out that the metrics for blurry areas on the slides and the coverslip give a high number of false positives when not analysed carefully.
More artifact detection results can be found in appendix A.

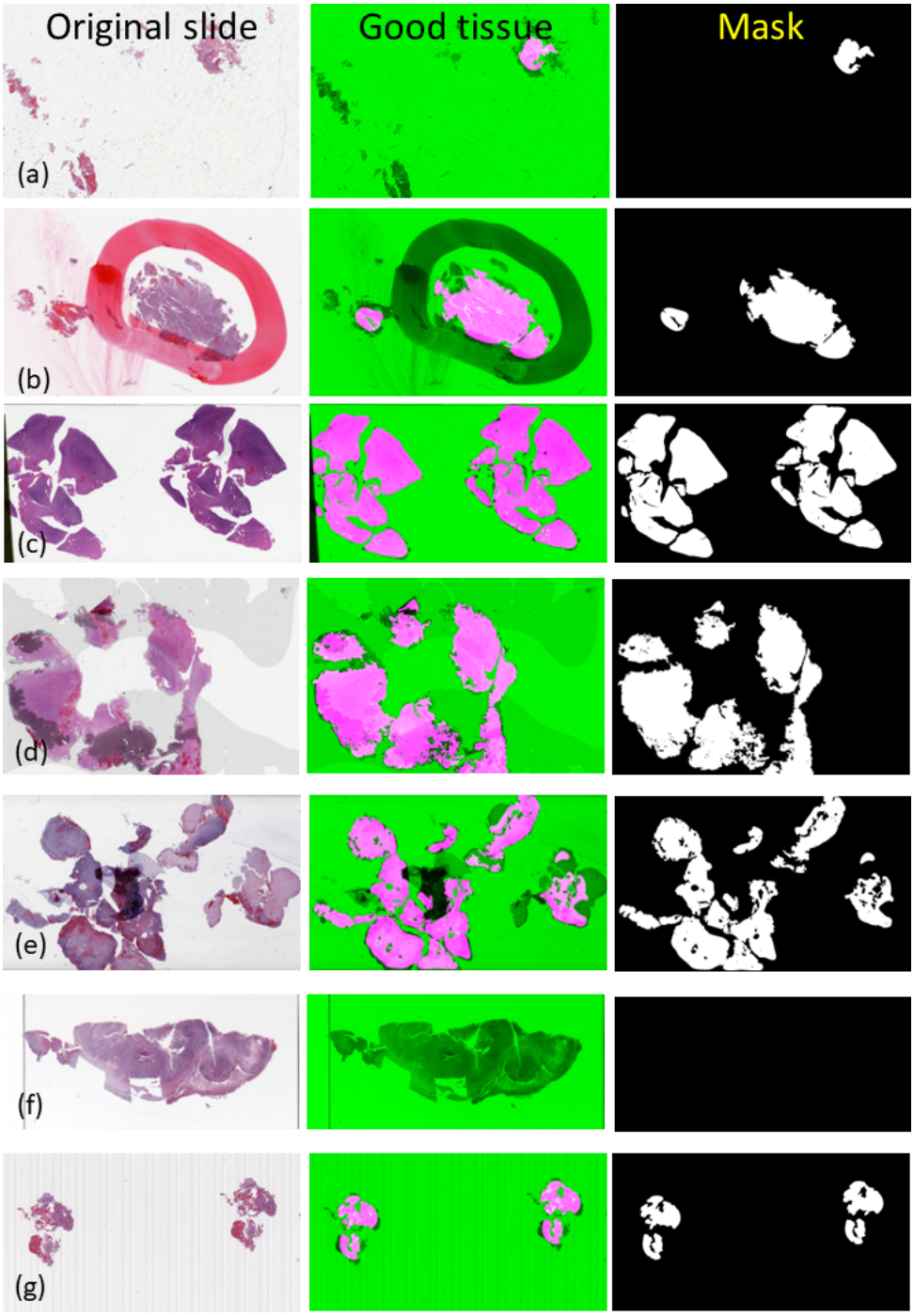

Figure 5.2: The results of the artifact detection in histoQC. On the left side the original slides are represented, in the middle the representation on which the pink tissue is the tissue that is suitable for analysis, in black/grey the artifacts detected and in green the background. On the right side the mask is shown that can be overlay over the original slide to extract the tissue suitable for analysis. Slide a-f represent the same slides as in figure 2.4, with (a) a dirty slide, (b) a slide with a pen marking, (c) a slide with on the left the cover slip, (d) a slide with air bubbles, (e) a slide with folded tissue, (f) a blurry slide and (g) a slide with grid noise in the background.

5.2 Similarity Analysis

Before starting the similarity analysis, the 78 slides on which histoQC did not detect any analysable tissue were removed from the dataset. Then the PCA was run as described in the methods section. The graphical representation of the PCA results can be found in figure 5.3.

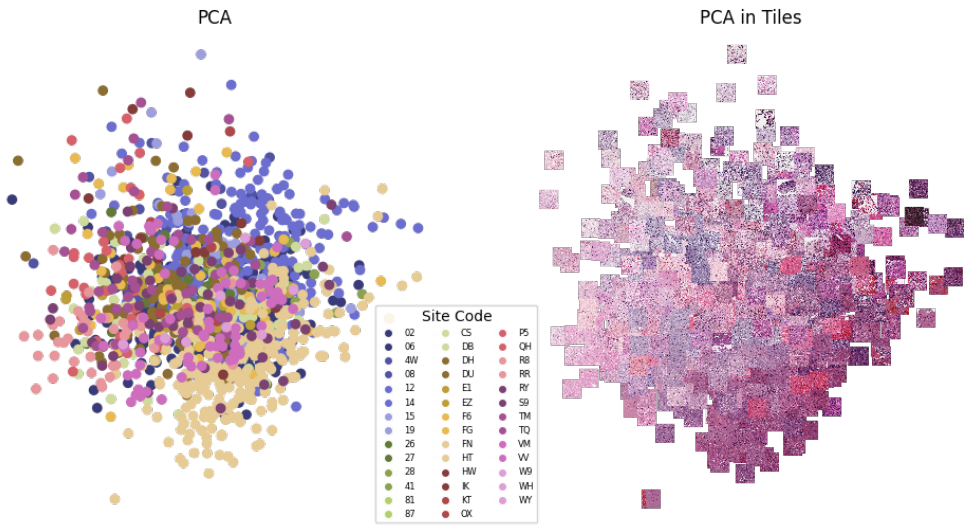

Figure 5.3: The results of the PCA analysis after deleting the slides on which no tissue was detected. On the left the points in the scatterplot are coloured based on the site at which the slide was created. On the right the points are replaced with a tile of the slides.

On the left side the PCA results are shown colour coded based on the site at which the slide is generated. It is visible that the slides that are generated on one site are more similar to each other than slides generated on different sites. For example the yellow dots representing site 'HT' are centred at the bottom of the point cloud.

On the right side the points in the PCA results are replaced with a tile of the original slides. In this visualisation, it is clear that the slides with a lower contrast are represented at the upper left part of the point cloud and the slides with a higher contrast are represented more at the bottom right of the point cloud.

Overall, there are no clear clusters in the PCA results, meaning that the

slides are more or less similar. However, even without clusters in the data, when looking at the PCA in thumbnails visualisation, one can see that from topleft to bottom right the tiles go from pale to dark coloured. So, it might be possible to classify the group into two based on the colour intensity in the slides.

Furthermore, when looking at figure 5.4 it is visible that some hospitals are more clustered at a distinct location in the point cloud than others. On the left side, hospitals '12' and 'HT' are coloured and as shown these hospitals are located at the border of the point cloud and more or less seperated from the other hospitals. On the right side, hospitals '06' and 'DU' are coloured and this representation shows that these two hospitals are more located in the center of the point cloud and are evenly distributed over it. This figure shows that the hospital at which a slide is generated, can influence the similarity between the slides.

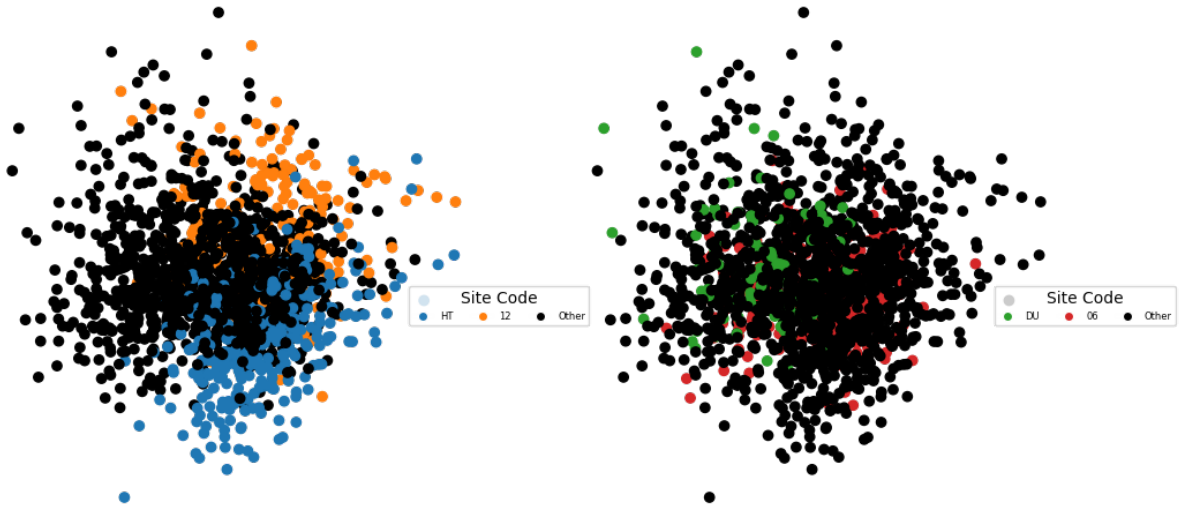

Figure 5.4: PCA results on the slides. On the left the hospital 'HT' is coloured blue and hospital '12' is coloured orange. The slides these hospitals produced are more similar to the other slides these hospitals produced than to the slides produced at other hospitals in the dataset. On the right hospital 'DU' is coloured green and hospital '06' is coloured red. The slides generated at these hospitals are in the middle of the point cloud and do not show any separation. The slides produced at hospital 'DU' and '06' are more representative for the slides from all hospitals.

Most explainable variables	
PC1	PC2
template3_MSE_hist chan2_brightness grayscale_brightness chan1_brightness_YUV deconv_c1_std	template4_MSE_hist template2_MSE_hist small_tissue_filled_percent fatlike_tissue_removed_percent fatlike_tissue_removed_num_- regions

Table 5.2: The five most explainable variables for PC1 and PC2

Next to this visual representation of the PCA results, the variance explained by principle component (PC) 1 and 2 are calculated. PC1 can explain 24.3% of the variance in the slides and PC2 19.6%, meaning that together PC1 and PC2 can explain 43.9% of the variance in the slides. At last the most explainable variables for PC1 and PC2 are subtracted, to find out on which ground the similarity found can be explained. In table 5.2 the most explainable variables for PC1 and PC2 can be found. This table represents that colour and brightness are the main explainable variables for PC1, since MSE_hist compares the colours in the slide with a template, and chan2_-brightness, grayscale_brightness and chan1_birhtness_YUV are measures for the brightness in the slide. For PC2 the most explainable variables are also colour based, since the first two explainable variables are both a comparison of the slide to a template. Next to this colour based explainable variables PC2 also has explainable variables based on what is in the slide, such as fatlike tissue and small tissue areas. A full explanation of all variables in the metrics output can be found in the appendix C.

To get an even better idea of the similarity between the slides, the PCA results are used as input for an UMAP analysis. This is done to get an idea of the similarity between the slides on all dimensions, where the PCA results explained above only represent two dimensions. The UMAP results can be found in figure 5.5. In this visualisation again on the left the points in the visualisation are coloured based on the site at which the slide was generated and on the left the points are replaced by a tile per slide. In these results the hospitals are more evenly distributed over the whole point cloud then in the PCA results. Furthermore, there are no distinct clusters in the UMAP

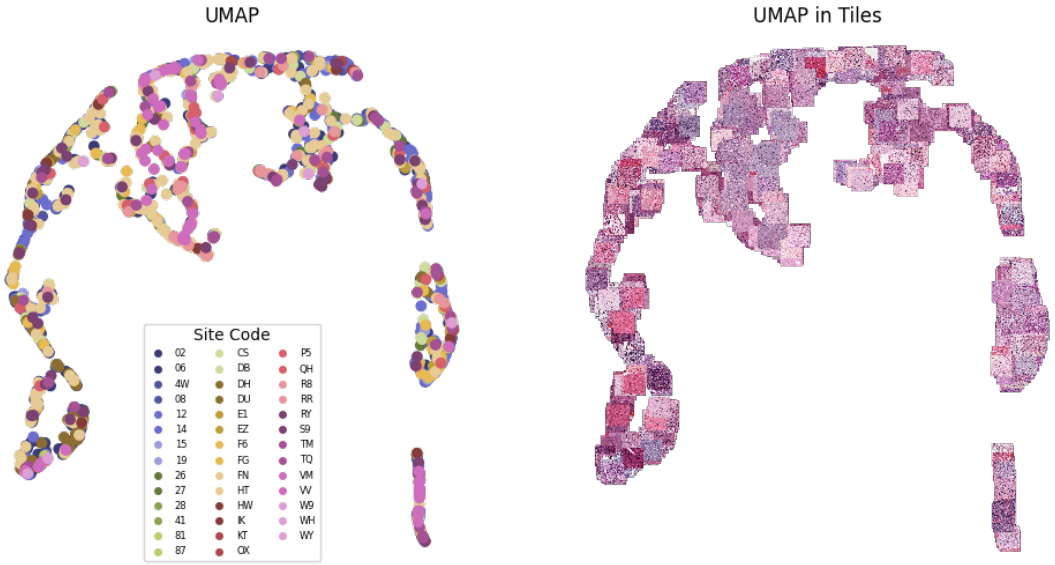

Figure 5.5: The results of the UMAP analysis with the PCA results as input. On the left the points are colour coded based on the site at which the slide is generated, on the right the points are replaced with a tile of the slide.

results. In the tile representation we see that the distinct separation of paler and darker slides is less significant in the UMAP analysis. There is however still a big variation within the slides, as the distance between the two tails is relatively large. This indicates that some slides are more alike than others. Colour does not seem the biggest factor to explain this variation based on this results.

6. Conclusion

The aim of this research paper is to assess histoQC as a quality control tool for the TCGA LGG and GBM dataset. In order to do this, the first research question was: 'Can histoQC accurately detect artifacts on the WSI in the TCGA glioma dataset?' The results show that histoQC in the default configuration was not able to accurately detect the artifacts on the slides. First histoQC did not report about dirt, air bubbles, folded tissue and grid noise, although we are interested in these artifacts. Second, the results on the amount of pen markings, cover slip and blurry areas were not accurate when compared to eyeballing. In order to increase histoQC's accuracy on the artifact detection, the configuration tool can be adjusted. For example, modules detecting dirt, air bubbles, folded tissue and grid noise can be added as classification modules by creating examples and masks that can be used to train a randomforest classifier. Other options to detect these artifacts are to add specific histoQC modules to the configuration pipeline to detect the artifacts, for example for the air bubbles the module bubbleRegionByRegion.py can be used to detect the air bubbles. For increasing the accuracy on detecting the pen markings and cover slip it would have been good to write down which slides contain these artifacts when manually checking all slides. This way examples from the dataset can be used to train the classification module to increase the accuracy. Furthermore, now all slides were manually checked to assess if there were artifacts present, statistically only a sample would have been enough to check the accuracy of histoQC. More precise, if the amount of a certain artifact detected by histoQC is 5%, it will be enough to check only 0.05 * 1704 = 85 slides manually to assess the accuracy of the results.

The second research question was: 'Can the output metrics of histoQC be used for performing a similarity analysis on the WSI in the TCGA glioma dataset?'. The results show that the histoQC metrics are a good source to

perform similarity analysis on the dataset. In order to further improve these similarity analysis it will be good to take a look at what metrics are used. The general metrics about the images were already removed, but currently the results on the artifacts that are detected are still included, however I am not interested in the similarity based on the artifacts, because this will not change robustness of trained models in the future. At last, it will be good to dive deeper into the UMAP results to find out what feature will explain the similarity and dissimilarity between the slides. Some options can easily be checked, for example if the type of tumour is relevant for the pattern in UMAP, can be checked by colour coding the points based on the type of brain tumour.

For further research it will be interesting to adjust the configuration pipeline in order to get it more specified on the artifacts or metrics of interest. For example more classifiers can be added in order to detect the other artifacts histoQC was not able to detect in the default configuration. Next to that the classifiers could be trained on templates chosen from the original dataset in order to increase the accuracy. Next to examples from the original dataset as templates also multiple examples can be provided to histoQC in order to train a more robust classifier.
Another interesting point for further research is using the masks histoQC creates to overlay the original slides to subtract the tissue that histoQC detected as suitable for further analysis. If only these high quality tissues of the slides are used for model training, it might result in better model performance.
Further improvement in model performance might be gained by using the similarity analysis results to divide the dataset in multiple clusters of less similar slides, so that the slides represent the over-all variability better. This way models will perform better on new hospital data, due to the models being trained on high variability data and therefor learns to ignore for example the colour differences in the slides, resulting in less biased models.
When the division in clusters based on the similarity analysis is proven to result in more robust models trained on the slide, one can even look into the similarity results as a basis for predicting which slides will be good for

robust model building.
At last, it will be interesting to check if the results of the similarity analysis change after colour-normalising the slides. This way the effect of colour normalisation on model building can be researched.

Acknowledgements

I want to thank the Norwegian Center for Molecular Medicine for making it possible to do my thesis with them. More specifically, I want to thank Sebastian Waszak and Birgit Kriener, for their help with the project and guiding me in the right direction. I really appreciate their time and support, which helped me to find the right analysis and give me insights in why I performed them.
Furthermore, I want to thank Artem Kaznatcheev for guiding me through the writing process. The feedback I received really helped me structure my writing and improved my paper.
At last, I want to thank my friends and family, for supporting me in the process and listening to my complaints and struggles over and over again.

Appendices

A. Artifact Detection

In the image in figure A.1 all the slides detected as slides with penmarkings are shown. When taking a look at all 57 of these slides, it is noted that not all of these slides actually contain penmarkings on them. Most likely, histoQC detects penmarkings in them due to the bright (red) colour in them on some of the places.

In the image in figure A.2 an example image of histoQC detecting a blurry image is shown, with its tile next to it. Altough histoQC detects blurriness in the whole slide, the tile shows that there is tissue on the slide that is not out of focus.

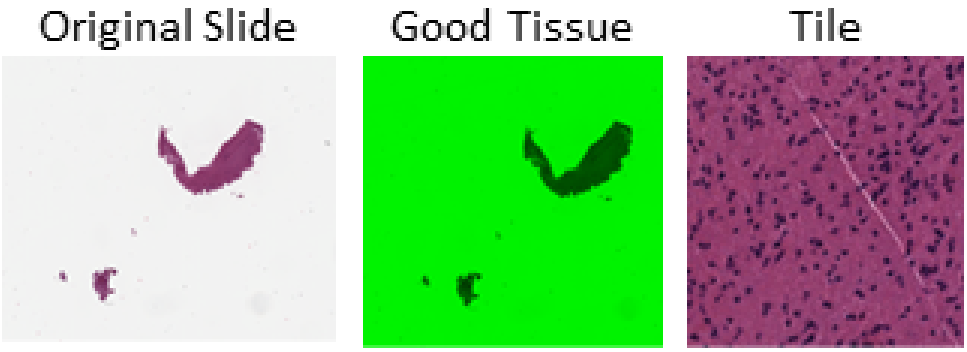

Figure A.2: Example image in which HistoQC found no analysable tissue because of the whole image being out of focus, however the tile created from the WSI being in focus.

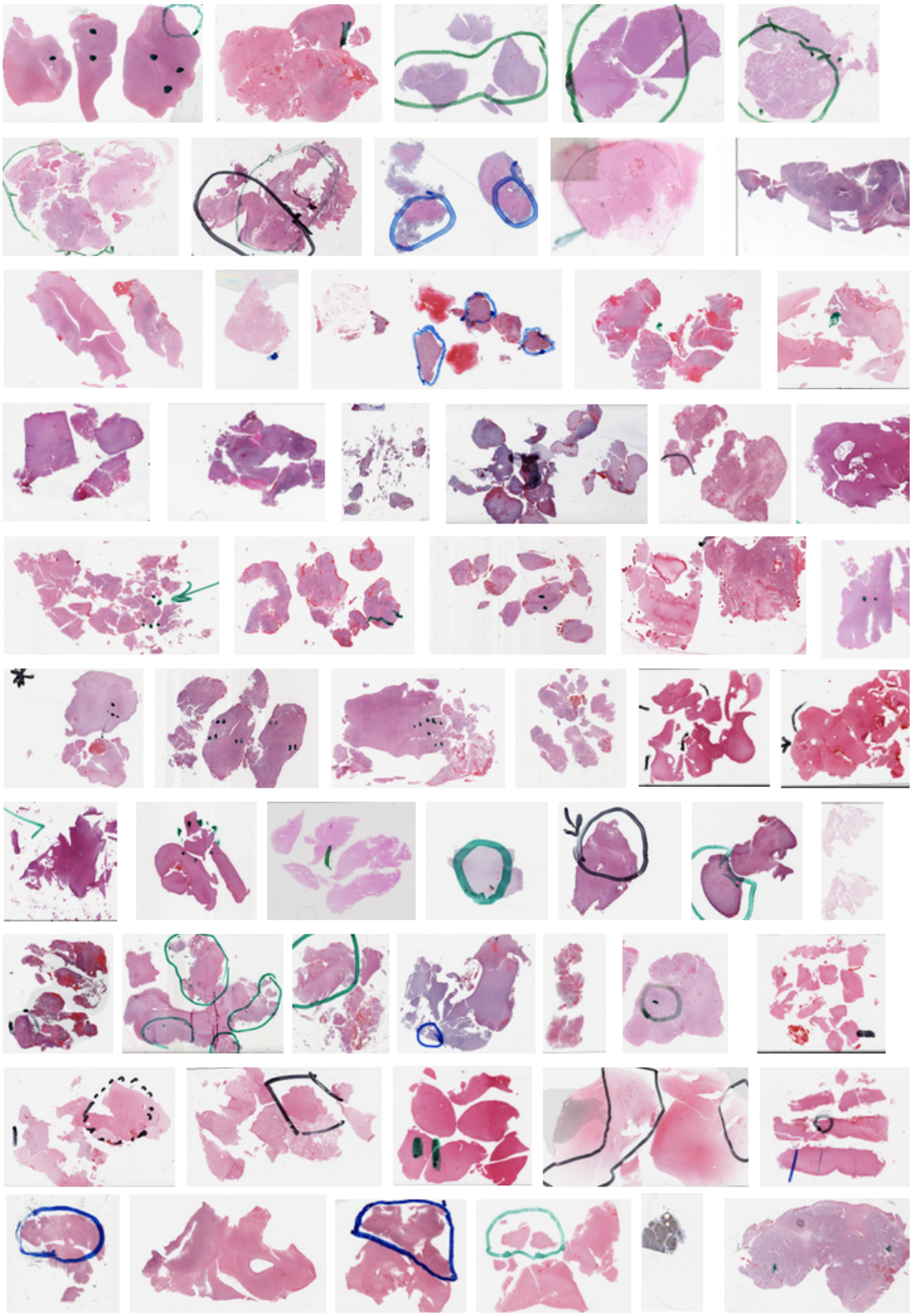

Figure A.1: All 57 slides histoQC detected to have penmarkings on them.

B. Complete Results

In this appendix I will show the complete similarity analysis results, also the PCA and UMAP results with the empty slides included.

B.1 All slides included

At first the variation in the variables of the output metrics of histoQC was assessed. The boxplot with this results are shown in figure B.1. Over all 51 metrics that are considered for the similarity analysis, the normalized variation is more or less the same. After analysing the boxplot and the variance

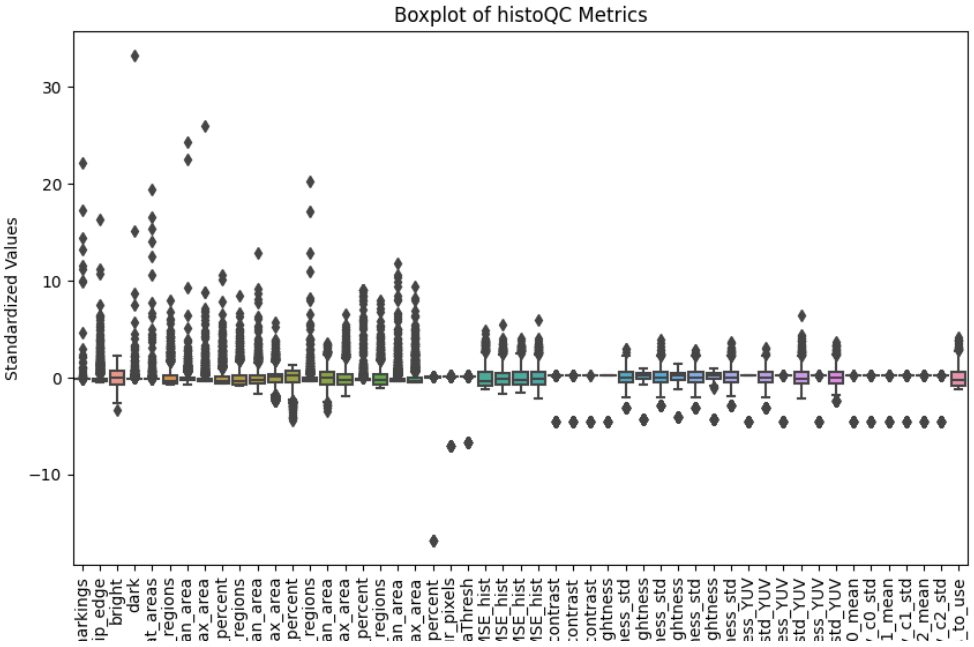

Figure B.1: The boxplot displaying the variance in the variables used from the output metrics of histoQC.

in the metrics, the PCA was performed to look for similarity between the slides. The results of this initial PCA analysis can be found in B.2. In this graph we can see that there are two groups in the data. When looking at the

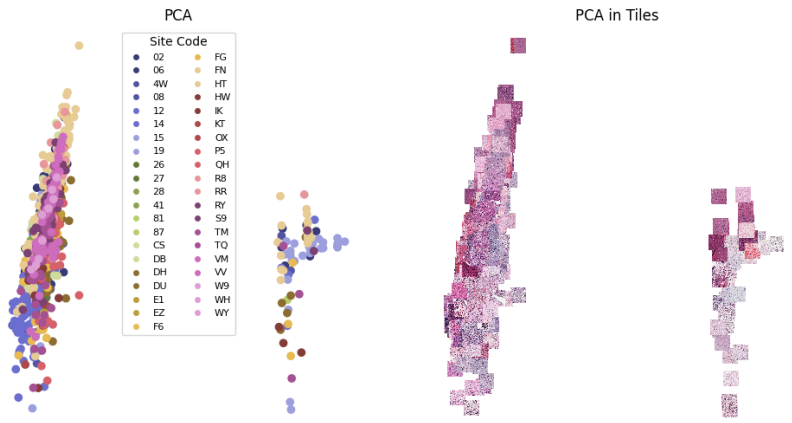

Figure B.2: The PCA results of all data. On the left the points are colour coded based on the hospital the slide was generated, on the right the points are replaced by the thumbnails of the slides. Two distinct groups are present in the data, with the left group represented with less pale slides than the right group. Further analysis indicated that the right group consists of the slides on which histoQC was not able to detect tissue.

PCA results in thumbnails, we can see that the group on the right contains paler images than the group on the left. The right group in the PCA results contains 79 slides, from which all 78 slides that contain no tissue. Therefore, for better representation of the results, these 78 slides were removed from the dataset and the PCA is then performed only on the slides on which histoQC did detect tissue. In this initial analysis PC1 explains 40.7% of the variance in the data and PC2 explains 12.2% of the variance. The most explainable variables can be found in table B.1.

Most explainable variables	
PC1	PC2
michelson contrast	template4_MSE_hist
rms contrast	small_tissue_filled_percent
chan3_brightness_YUV	fatlike_tissue_removed_num_regions
deconv_c2_std	template2_MSE_hist
deconv_c1_std	fatlike_tissue_removed_percent

Table B.1: The five most explainable variables for PC1 and PC2 in the analysis including all slides

B.2 Only slides with tissue

After removing the slides on which histoQC was not able to detect any tissue, the variance in the metrics was re-evaluated. If we look at the variance in the metrics shown in the boxplot in figure B.3 and compare it with the one in figure B.1 it is clear that when removing the slides with no tissue, results in less negative variance. This can be explained by the fact that histoQC assigns -100 to all values when it is not able to detect any tissue on the slide.

The results after removing the empty slides are discussed in the results section.

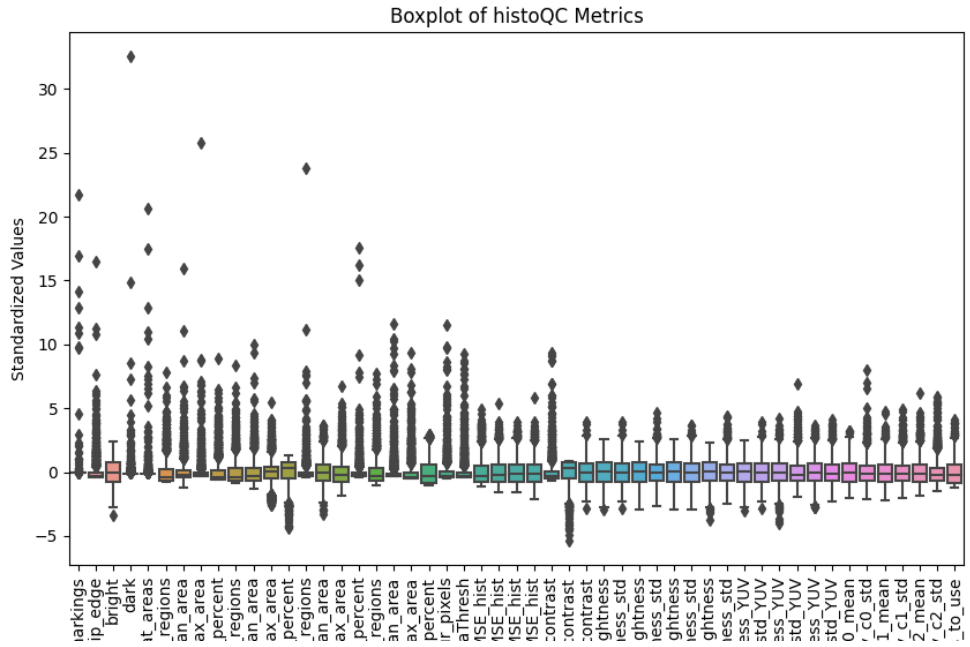

Figure B.3: The boxplot displaying the variance in the variables used from the output metrics of histoQC.

C. HistoQC Output metrics

Metric	Description
filename	The name of the slide used as input
image_bounding_box	The rectangular area that encompasses the tissue region on the slide
base_mag	The magnification used to digitize the WSI
type	The type of scanner used to digitize the WSI
levels	
height	The height of the slide in pixels
width	The width of the slide in pixels
mpp_x	The magnificaiton per pixel in the x-direction
mpp_y	The magnification per pixel in the y-direction
comment	The details about the scanning process
pen_markings	The percentage of the slide that is covered by pen markings
coverslip_edge	The percentage of the slide that contains a cover slip.
bright	The brightness of the slide, represents if the whole slide is in the same brightness
dark	The darkness of the slide, represents the level of darkness or contrast on the slide
flat_areas	The percentage of the slide that is flat or has low curvature
fatlike_tissue_removed_num_regions	The number of regions containing fat-like tissue that are removed
fatlike_tissue_removed_mean_area	The average area of fat-like tissue areas that are removed

fatlike_tissue_removed_max_area	The maximum area of fat-like tissue that is removed from the slide
fatlike_tissue_removed_percent	The percentage of fat-like tissue that is removed from the slide
small_tissue_filled_num_regions	The number of regions in the slide that contain small tissue sections
small_tissue_filled_mean_area	The mean area of the regions in the slide that contain small tissue sections
small_tissue_filled_max_area	The largest area of the regions in the slide that contain small tissue sections
small_tissue_filled_percent	The percentage of the slide that is covered by small tissue sections
small_tissue_removed_num_regions	The number of regions containing small tissue sections that are removed
small_tissue_removed_mean_area	The average area of regions containing small tissue sections that are removed
small_tissue_removed_max_area	The largest area of all regions containing small tissue sections that is removed
small_tissue_removed_percent	The percentage of the slide that contains small tissue sections that are removed
blurry_removed_num_regions	The number of regions containing blurry sections that are removed
blurry_removed_mean_area	The average area of blurry sections that is removed
blurry_num_max_area	The largest blurry section that is removed
blurry_removed_percent	The percentage of the slide that contained blurry sections that are removed
spur_pixels	The percentage of the slide that contains small, isolated pixels that are not connected to the main tissue regions

areaThresh	The percentage of the slide that is above a certain threshold for tissue area. If this value is low it means that there is a low density of tissue regions.
template1_MSE_hist	Compares the colours in the slide to the colours in template 1
template2_MSE_hist	Compares the colours in the slide to the colours in template 2
template3_MSE_hist	Compares the colours in the slide to the colours in template 3
template4_MSE_hist	Compares the colours in the slide to the colours in template 4
tenenGrad_contrast	Represents the contrast on the slide.
michelson_contrast	A measure for contrast that takes into account the maximum and minimum brightness values of the slide. It is calculated as the difference between the maximum and minimum brightness values divided by the sum of the maximum and minimum brightness values.
rms_contrast	A measure for the contrast that is calculated by the root mean square difference between the brightness values in the slide.
grayscale_brightness	A measure for the brightness of the grey colours in the slide, represents its overall lightness or darkness.
grayscale_brightness_std	The standard deviation in the grayscale brightness.
chan1_brightness	The brightness of the red colours in the slide.
chan1_brightness_std	The standard deviation in the brightness of red colours.
chan2_brightness	The brightness of the green colours in the slide.
chan2_brightness_std	The standard deviation in the brightness of green colours.

chan3_brightness	The brightness of the blue colours in the slide.
chan3_brightness_std	The standard deviation in the brightness of blue colours.
chan1_brightness_-YUV	The mean channel brightness of the red color channel of the slide after converting to YUV color space.
chan1_brightness_-std_YUV	The standard deviation in chan1_brightness_YUV.
chan2_brightness_-YUV	The mean channel brightness of the green color channel of the slide after converting to YUV color space.
chan2_brightness_-std_YUV	The standard deviation in chan2_brightness_YUV.
chan3_brightness_-YUV	The mean channel brightness of the blue color channel of the slide after converting to YUV color space.
chan3_brightness_-std_YUV	The standard deviation in chan3_brightness_YUV.
deconv_c0_mean	The mean deconvolution on channel 0.
deconv_c0_std	The standard deviation in deconvolution on channel 0.
deconv_c1_mean	The mean deconvolution on channel 1.
deconv_c1_std	The standard deviation in deconvolution on channel 1.
deconv_c2_mean	The mean deconvolution on channel 2.
deconv_c2_std	The standard deviation in deconvolution on channel 2.
pixels_to_use	The number of pixels in the slide that are used for analysis.

Table C.1: The output metrics histoQC generates with the default configuration file.

D. HistoQC Output metrics

In the following table all modules available in histoQC are described including the operations. This table is based on the table in the supplemental material of the journal of clinical oncology, histoqc: An open-source quality control tool for digital pathology slides. [7]

File module	Operation	Description
MorphologyModule.py	removeSmallObjects	Remove small items from the image. This is typically done for reducing small pixel noise, dust, etc.
	fillSmallHoles	Fill in small/medium sized "holes" in images.For example, lumen spaces in tubules often are detected as background and removed from the final mask. This module will fill them in.
LightDarkModule.py	getIntensityThresholdOtsu	Thresholds the image based on dunamic Otsu threshold.
	getIntensityThresholdPercent	Thresholds the image based on user supplied values. This is good for detecting where the tissue is on the slide (non-white) and where folded tissue may be (very dark).
HistogramModule.py	getHistogram	Makes a histogram image in RGB space.
	compareToTemplates	Compares the image's histogram to template images provided by the user.
DeconvolutionModule.py	seperateStains	Performs stain deconvolution using skimage's built in matrices.
ClassificationModule.py	pixelWise	Applies an RGB based classifier to the image whose values come from a user inputted TSV.
	byExampleWithFeatures	Computes features of template images provided by the user which have associated binary masks indicating positive and negative classes. Trained classifier is then used on images. Excellent for e.g. pen detection (with texture), cracks, etc.
BubbleRegionByRegion.py	roiWise	Detect contours of lines of air bubbles on slide. Contains exemplar of how to use HistoQC to iteratively loop over very large images at high magnitude. (still work in process)
BrightContrastModule.py	getBrightnessGray	Computes the average value of the image in gray colour space, which ultimately represents how bright the image is perceived.
	getBrightnessByChannelinColorSpace	Computes a triplet (one per colour channel) in the desired colour space. Useful for detecting outliers.
	getContrast	Computes both RMS and Michelson contrast metrics.
PenMarkingModule.py	identifyPenMarkings	Identifies pen markings on a pixel by pixel basis by using user supplied TSV file of colour values. This is usually suitable when the marking is very different from the staining (e.g. green/blue marker on pink tissue).
BlurDetectionModule.py	identifyBlurryRegions	Uses a Laplace matrix to determine which regions in the image are likely blurry.
BasicModule.py	getBasicStats	Pulls out metadata from image header.
	getMag	Pulls out base magnification. This is required by histoQC. In the future we'll add ability to predict magnification.
	finalComputations	Computes the final number of pixels available in the output image. Too high or low of a number often indicate incorrect processing or image outliers.
	finalProcessingSpur	Removes spurious morphology from the final mask. Essentially small "arms" of tissue are rounded off and removed.

	finalProcessingArea	Removes larger islands from the output masks, e.g. isolated pieces of tissue.
saveModule.py	saveFinalMask	Saves both the output mask from HistoQC but also the overlay on the original image.
	saveThumbnails	Save thumbnails for easier viewing. This needs to be completed for the UI to work.

Table D.1: Modules available in histoQC including their operations and discription of the operations.

Bibliography

[1] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep learning in cancer pathology: A new generation of clinical biomarkers," *British journal of cancer*, vol. 124, no. 4, pp. 686–696, 2021.

[2] A. Shmatko, N. Ghaffari Laleh, M. Gerstung, and J. N. Kather, "Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology," *Nature cancer*, vol. 3, no. 9, pp. 1026–1038, 2022.

[3] J. Van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: The path to the clinic," *Nature medicine*, vol. 27, no. 5, pp. 775–784, 2021.

[4] G. Smit, F. Ciompi, M. Cigéhn, A. Bodén, J. van der Laak, and C. Mercan, "Quality control of whole-slide images through multi-class semantic segmentation of artifacts," in *Medical Imaging with Deep Learning*, 2021.

[5] Y. Chen, J. Zee, A. Smith, *et al.*, "Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies," *The Journal of pathology*, vol. 253, no. 3, pp. 268–278, 2021.

[6] F. M. Howard, J. Dolezal, S. Kochanny, *et al.*, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nature communications*, vol. 12, no. 1, p. 4423, 2021.

[7] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, "Histoqc: An open-source quality control tool for digital pathology slides," *JCO clinical cancer informatics*, vol. 3, pp. 1–7, 2019.

[8] S. Lapointe, A. Perry, and N. A. Butowski, "Primary brain tumours in adults," *The Lancet*, vol. 392, no. 10145, pp. 432–446, 2018.

[9] P. A. T. E. Board, "Adult central nervous system tumors treatment (pdq®)," in *PDQ Cancer Information Summaries [Internet]*, National Cancer Institute (US), 2022.

[10] A. N. Mamelak and D. B. Jacoby, "Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (tm-601)," *Expert opinion on drug delivery*, vol. 4, no. 2, pp. 175–186, 2007.

[11] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[12] R. D. Fields, A. Araque, H. Johansen-Berg, *et al.*, "Glial biology in learning and cognition," *The neuroscientist*, vol. 20, no. 5, pp. 426–431, 2014.

[13] M. Kapoor and V. Gupta, "Astrocytoma," in *StatPearls [Internet]*, StatPearls Publishing, 2021.

[14] M. J. Van den Bent, M. Reni, G. Gatta, and C. Vecht, "Oligodendroglioma," *Critical reviews in oncology/hematology*, vol. 66, no. 3, pp. 262–272, 2008.

[15] F. E. Bleeker, R. J. Molenaar, and S. Leenstra, "Recent advances in the molecular understanding of glioblastoma," *Journal of neuro-oncology*, vol. 108, pp. 11–27, 2012.

[16] A. C. Tan, D. M. Ashley, G. Y. López, M. Malinzak, H. S. Friedman, and M. Khasraw, "Management of glioblastoma: State of the art and future directions," *CA: a cancer journal for clinicians*, vol. 70, no. 4, pp. 299–312, 2020.

[17] B. Tran and M. Rosenthal, "Survival comparison between glioblastoma multiforme and other incurable cancers," *Journal of Clinical Neuroscience*, vol. 17, no. 4, pp. 417–421, 2010.

[18] H. Zong, R. G. Verhaak, and P. Canoll, "The cellular origin for malignant glioma and prospects for clinical advancements," *Expert review of molecular diagnostics*, vol. 12, no. 4, pp. 383–394, 2012.

[19] H. Zong, L. F. Parada, and S. J. Baker, "Cell of origin for malignant gliomas and its implication in therapeutic development," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 5, a020610, 2015.

[20] N. Elazab, W. GabAllah, and M. Elmogy, "Computer-aided diagnosis system for grading brain tumor using histopathology images based on color and texture features," 2022.

[21] C. for Cancer Genomics, *The cancer genome atlas program*, `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`, Accessed: 2023-06-14.

[22] A. Janowczyk, *Histoqc*, `https://github.com/choosehappy/HistoQC`, 2019.