Automated HER2 Scoring of Breast Cancer Tissue using Upconverting Nanoparticle Images

Master Thesis at Lumito AB

Adam Belfrage & Alexander Wik

LUNDS UNIVERSITET

LUMITO

Centre for Mathematical Sciences
Lunds Tekniska Högskola
Sweden

Abstract

Computer aided pathology is becoming more and more of a requirement within pathology due to increased demand of individualised treatments and personalised medicine. Because of the advance of digital pathology in recent years, where a high resolution camera acquire images of microscope slides, pathologists can now assess tissue samples in digital images. This has enabled automatic assessment of pathological images. A specific area of interest is the quantification of HER2-receptors in breast cancer tissue which decides if targeted therapy can be used or not. The standard staining within the area obstructs cell morphology and the images are difficult to analyse and classify. Existing automated HER2-classification methods in the field rely heavily on colour consistency or are neural networks which are difficult to interpret. Lumito AB has developed a reagent kit that, via laser and upconverting nanoparticles, demonstrates the HER2-expression in separate images that does not interfere with cell-morphology. These images are potentially more suitable for traditional image analysis and could potentially enable the possibility to develop simple, fast and interpretable algorithms that could quantify the HER2-expression and classify tissue samples. In this project, two algorithms were developed for classification of the upconverting nanoparticle based images. They were considered to be simple in the sense that the bases of classifications would be easy to explain to a pathologist due to the fact that they were inspired by the guidelines that pathologists use for HER2-classification. The algorithms performed on par with a pathologist and could be used as a screening tool, reducing the pathologist's workload. The algorithms were also accurate in classification of the HER2 positive and equivocal tissue samples but fail to classify these unambiguously, and a pathologist would still have to assess these samples manually. It is difficult to say how well the algorithms performed in reality due to the relatively small data-set. This project should be seen as a proof of concept and future work would have to be done to further validate and improve the results even though the start is promising.

Contents

Preface

This master thesis project has been conducted at Lumito AB from the middle of January to the beginning of June. The purpose of the project was to evaluate the possibility of automated image classification of breast cancer types using images derived with the Lumito technology. We would like to give a special thanks to our supervisors Ida Arvidsson at the Mathematical deparment of LTH and Klas Bergren at Lumito AB. We would also like to thank Magnus Helgstrand, Marcus Valtonen Örnhag, Mathias Mickert and David Belfrage at Lumtio AB that also contributed to the project and helped it moving forward. Finally we would like to thank all the other Lumito staff that made our time at the company rewarding and fun.

1 Introduction

Throughout the years, the number of tissue samples and demands of individualised treatments for breast cancer have increased, which creates a high workload for the pathologists. This, in combination with the development of digital glass-slide scanning over the last few years resulted in the area now known as digital pathology. Instead of physically looking on a tissue sample through a microscope, a camera with high resolution acquires an image of the tissue sample that could be displayed on a computer screen. This brought a numerous simplifications in contrast to traditional pathology practice such as sharing information[15]. However, the pathologist still needs to asses each image carefully and thoroughly which takes time. When the amount of tissue samples increases the assessment time becomes the bottleneck. That is where computer aided diagnostics entered the picture. Similar thing have been done in numerous fields within healthcare already and in pathology it is very much on its way [19] [20]. Within computer aided pathology and other fields of computer aided diagnostics interpretability is of key importance. Interpretability is 'the capacity to provide explanations that are relevant and interpretable by experts in the field' [21]. The intepretability for a diagnostic tool is important in order to gain confidence from the practitioners within the field [21].

A typical pathological assessment is to diagnose different cancer types. Different cancer types have different treatment plans, therefore it is important to classify them correctly. Concerning breast cancer one can classify different cancer types based on a protein expression called Human epidermal growth factor 2 receptor(HER2) [11]. The protein expression of HER2 can be visualised by immunohistochemistry. With this technique a pathologist can thereby make an assessment if there are an abnormal amount of HER2 proteins in the tissue or not. However, immunohistochemistry obscures the cell morphology and colour variations can occur dependent on tissue staining and detection systems. By using upconverting nanoparticles(UCNP) to obtain the images, a separate greyscale image that only contains the expression can be created. The expression separation as well as the non-obstruction of cell-morphology are properties which potentially could make these images advantageous in interpretable automated HER2-scoring compared to traditional immunohistochemistry images.

1.1 Purpose of the Thesis

The purpose behind this project is to investigate the possibility of classifying pathological images of breast cancer tissue obtained with UCNP:s into different HER2-cancer categories using interpretable image analysis methods. The main motivation behind the project is that the novel images obtained with UCNP:s have many desired qualities for image analysis. It would be interesting to compare classification performance using UCNP-images to existing methods, developed on traditional pathological images, in the field. If they can be classified with sufficiently high accuracy by using interpretable methods then much is won as interpretability is held in high regard within the medical field.

1.2 Limitations

The main scope will be focused on the development of simple image analysis methods for classification as an alternative to neural network classification and previous classification methods. It is not an objective to out-perform existing methods in the terms of accuracy, the focus is on interpretability. This report will only focus on classification on HER2-scores in breast cancer and not any other classification, meaning that more (or less) success could be achieved in another field, studying a different type of tissue. Considering HER2-evaluation, only the guidelines of protein overexpression were used while implementing and evaluating the performance of the image analysis methods.

1.3 Report framework

Firstly some background knowledge to familiarise the reader with digital pathology and theory used for the proposed classification algorithms is presented. The methodology will be explained after and results follow after that. The results will be, for each classification algorithm developed, presented in statistical metrics such as confusion matrices but also in a graphical way such as plots and images. The result section will be followed by a discussion section, where the results will be discussed and evaluated. Finally, conclusions from the thesis will be presented. There will also be a short paragraph discussing ethical aspects.

2 Background

2.1 Breast cancer and Human Epidermal Growth Factor 2 Receptor

The HER2 gene creates receptors named HER2 2 proteins on the cell membranes. These proteins controls the growth factor of a cell. In certain types of breast cancer, approximately 12%-15% of all breast cancers, the HER2 gene creates an abnormal number of copies of itself [8]. The increased number of HER2 genes creates an increased number of HER2 receptors on the cell membrane. These proteins signals the cells to grow and divide in a faster pace than normal. An abnormal growth of cells will result in an even quicker expansion of the tumour and in that way, a more aggressive cancer [6].

By being a receptor on the outside of the membrane the HER2 protein can be used for targeted therapy [3]. HER2 positive patients can be treated with several different types of targeted therapy treatments. All targeted therapy treatments have in common that the drug finds the cancerous cells since it connects with the HER2 receptor on the outside of the cell. As an example, lab manufactured antibodies can be used to attach to the cancerous cell. The antibodies can hold another drug that either destroys the cell or inhibits the receptor to work properly [4]. The treatment plan will therefore partly depend on the HER2 expression in a tumour. Therefore the testing of HER2 expression is of key importance while conducting a diagnosis and setting up a treatment plan for patients diagnosed with breast cancer.

2.2 Traditional immunohistochemistry

Immunohistochemistry(IHC) is a method that can be used to find the HER2 expression of cells [6].The method is based on antibodies in a staining which binds to antigens on the cell membrane. The antibodies can then bind an enzyme which creates a precipitate that can be seen through a brightlight microscope. To complete this procedure there are a couple of preparation steps that must be completed before a proper image is created. Firstly, the antigens needs to be activated in the sample. In that way antibodies from the staining can connect to them, this is called antigen retrieval. Secondly, the sample must be stained with a blocking for non-specific antigens; since the antibody in the final staining only should bind to the specific antigen of interest. In HER2 evaluation, the antibody that connects to the HER2 protein is also connected to a secondary antibody which holds the enzyme horseradish peroxidase (HRP), the HRP oxidize 3,3-Diaminobenzidine (DAB)[18] [1]. When oxidised DAB creates a brown precipitate which can be seen through the brightlight micropscope.

2.3 ASCO-guidelines

While classifying HER2 expression, the ASCO-guidelines are used by pathologists to determine if the tumor is HER2-positive or HER2-negative [24]. If the tissue consists of an abnormal number of HER2 proteins it will be classified as HER2-positive, else it is classified as HER2-negative. As explained in the guidelines, these are not a set of rules, rather guidelines to create a more standardised way of classifying HER2 expressions. The evaluation can be completed by either finding protein overexpression or by gene amplification. According to the guideline for HER2 testing with imunohistochemistry, there are four different classes, two negative, one equivocal and one positive. If classified as equivocal, a new test must be completed, either with another IHC test or with another type of test. The two negative classes are named IHC 0 and IHC 1+, the equivocal is named IHC 2+ and the positive is named IHC 3+. However, IHC is not used in this project, therefore the classes will be called HER2 instead of IHC. A simplification of the decision tree that is used to evaluate the tumour, according to the ASCO-guidelines, is explained in Figure 1. The tree is not completely mutually exclusive or collectively exhaustive, and special cases can occur. These cases occur seldom and are explained further in the ASCO-guidelines.

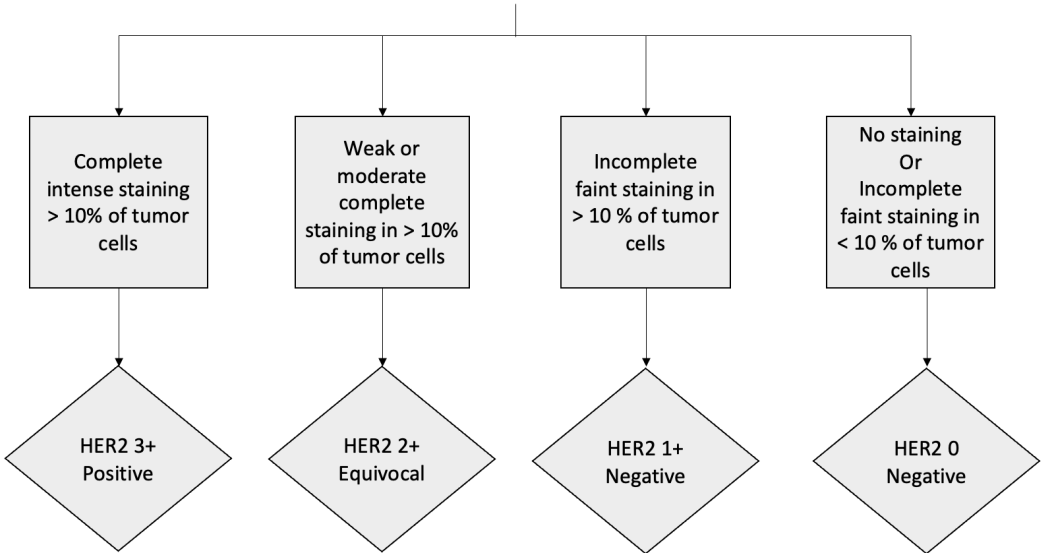


Figure 1: ASCO-guidelines for evaluation of tissue samples using IHC

2.4 Whole slide imaging

Whole slide imaging (WSI) is a term that is used within digital pathology and describes the image of an entire slide with a tissue sample. Each WSI consists of smaller images (tiles) acquired by a camera with microscopic resolution.

A full tissue sample can consist of several hundreds of tiles and occasionally over a thousand, which creates the need for stitching. Stitching algorithms calculates optimal tile translations to combine the tiles to make up the WSI without visible defects (such as edges or white gaps). This is typically a challenge since the spatial position of tiles needs to be derived, and concurrently finding the correct overlap between adjacent tiles. Small errors can accumulate to greater errors in the end of the stitching chain. Stitching follows three steps:

1. Compute candidate translations between adjacent tiles. Meaning that each tile is shifted spatially (in x and y directions) with respect to the adjacent tiles.
2. Adjust said translations to reduce errors in the stitched image.
3. Compose the constructed WSI based on the stitched images.

For step 1 there are two ways to calculate an appropriate translation - feature based[9] or Fourier based [12]. Feature based calculations use feature extraction to find matching features in two adjacent tiles and then use these features to calculate an overlap. Fourier based calculations use the Fourier transform to extract frequency features to calculate the overlap. If you crop out 100 edge pixels per side in every tile you can extract the frequency image of the cropped out parts for two adjacent tiles (for the side where they align). Most likely the frequency content would be very similar but the phase slightly different and in that way you can calculate the appropriate translation. Step 2 is a mathematical optimisation, optimise over calculated translations with respect to an optimisation criteria. The final step is a way to connect all the stitched images to one another to make up the final WSI. The goal is to make each tile only connected once to the final image. For an image you would find a way to connect all tiles so that each tile is stitched to only one of its neighbouring tiles. This is done for the entire set of tiles to build up the WSI. You would then run into the problem to find the optimal stitching path that creates the best image. Algorithms like minimum spanning tree solves this. [10]

After the stitching a blending operation is applied on each tile so that the overlap between the tiles will be shown only once and correctly. This finally results in a WSI of the entire tissue sample in high resolution where the dimension can vary. Each tile is typically of the dimension 2048x2048x3 (where the 3 describes the colour channel) and since each WSI can consist of a different amount of tiles there is no fixed dimension. The WSI:s are stored in a pyramidal way. Pyramidal means that the WSI is stored in lower resolutions in increments. The images in lower resolutions have been interpolated from the baseline image which is in the highest resolution. This is done tile wise for the entire sample. This allows zooming in a sample much similar to what you would do with an analogue microscope. The image in 3-channels is typically referred to as a *brightfield* image. [13][10]

In a WSI a control product array(CPA) can be added. This is an array of groups of cultivated cells corresponding to different types of categories for the tissue sample that is being investigated. Typically, for investigating HER2-expression the CPA will consist of cell groups in pellet like structures corresponding to different HER2-categories. This is used for comparison to the actual tissue. If the tissue has expression that resembles one of the categories then it is more likely that it belongs to that category. The CPA is typically appended to the bottom or the side of the slide. A typical WSI is shown in Figure 2 with a CPA appended to the bottom. And as can be seen below a CPA is consisting of four pellet like structures.

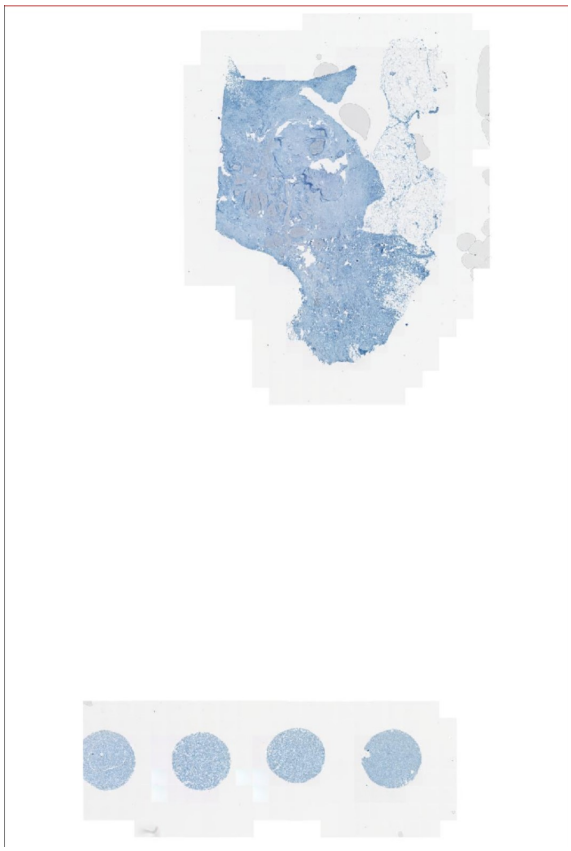


Figure 2: An example of a WSI with a CPA appended in the bottom

2.5 Upconverting nanoparticles

Lumito has generated pathological images on breastcancer tissue samples, using UCNP:s. UCNP:s make use of the anti-stokes effect to upconvert the energy of the light meaning that an illumination of these particles with longer wave-length light will result in emission of shorter wave-length light. These particles are bound to anti-bodies which in its turn binds to the protein expression of interest, similar to IHC. The anti-bodies are smeared on to a tissue sample which is illuminated with an infra-red laser. The particles will then emit light in the visible spectra around 400-700 nm. This emission is subsequently caught by a camera.

Since the particle free tissue not will emit any light the resulting image will be an image in greyscale where the bright parts corresponds to where the anti-bodies have bound to the protein-expression or receptor of interest. The particles are resistant to photobleaching and quenching (decrease in light intensity) but can occasionally demonstrate high numbers of non-specific binding (which means particles not specifically bound to the genetic expression of interest) which can have misleading consequences [14].

These UCNP-images are specifically interesting due to their simplicity. The fact that they present a high contrast between the protein-expression and other tissue in an almost binary sense could lead to a number of simplifications within the pathological area. Not only could it be easier for pathologist to asses and find protein-expressions of interest but it could also mean that the introduction of computer aided diagnostic tools could be facilitated. These tools could be quite simple and interpretable in contrast to such tools derived from neural networks. The interpretability of a computer aided diagnostic tool is held in high regard within medicine. If the basis of a decision or classification from a tool can be easily explained and derived from medical features then that tool is less likely to cause distrust and more probable to be applied [21]. Below a typical UCNP-image is shown with a brightfield image of the same sample appended to the right. Further down an comparison with traditional IHC-images is shown.

Figure 3: UCNP-image, to the left, where the bright parts corresponds to HER2-expression. A brightfield image to the right, displays the same tissue stained with hematoxylin

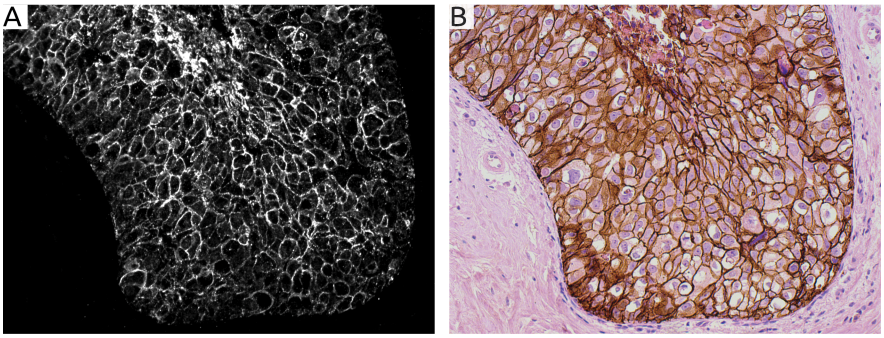


Figure 4: UCNP-image compared to an IHC-image

2.6 Bayesian classification

This section gives an overview of the theroy for Gaussian Naive Bayesian Classifiacations, and more details can be found in [17]. Gaussian Naive Bayesian Classification is based on Bayes theorem in combination with the naive assumption that conditional independence between features given the class holds. Bayes theorem describes the relation between posterior probability, likelihood, prior probability and evidence according to equation 1,

$$posterior = \frac{prior * likelihood}{evidence} \tag{1}$$

which can be expressed as the probability that a set of features would belong to a specific class, according to equation 2. For equation 2-4, let C denote the class, F denote feature, σ denote variance, μ denote mean, k denote which class and p denote teh probability.

$$p(C_k|F) = \frac{p(C_k) * p(F|C_k)}{p(F)} \tag{2}$$

While training the model, the mean value and variance is found given each class for all features while also calculating the probability for each class appearing. In that way, Gaussian distributions, each with the dimension of the number of features, can be found given each class, according to equation 3.

$$p(F_i|C = k) = \frac{1}{\sqrt{2 * \pi * \sigma_{i,k}}} * e^{\frac{1}{2*\sigma_{i,k}} * (F_i - \mu_{i,k})^2} \tag{3}$$

Prediction is performed by finding the C_k with the largest probability given the features, which is found by applying Bayes theorem. An iteration over every class is performed for each class the prior for that class is multiplied with the product of probabilities that each feature would appear given the class. After iterating over every class, the class which produces the maximum probability is chosen as the predicted class, see equation 4.

$$Class(F) = \text{argmax}_\text{k}(p(C|F)) = \text{argmax}_\text{k}(p(C_k) * \prod_i p(F_i|C = k)) \tag{4}$$

2.7 Singular value decomposition

Singular value decomposition (SVD) is a common way to reduce dimensionality of a data set by extracting the components of a data set that explains or contains the significant amount of variation within the set. The SVD could be seen as a simple factorisation over multiple dimensions. It is described mathematically as

$$X = U\Sigma V^T \tag{5}$$

a feature matrix X is factorised into three matrices. These corresponds to the right respective left eigenvectors of X multiplied by the eigenvalues. The matrices V and U will correspond to the principal directions whereas Σ will correspond to the principal components (or commonly referred to as principal scores). This means, that V and U explains the directions of where there is most variance and Σ will be projections on the these directions. This is an operation to find out if one could reduce the dimensionality, typically only a few principal directions explains most of the variation within the data set. One would then use only these directions to try to model the distribution that is being investigated. These directions can also be used to illustrate the variations within the high dimensional dataset. In this report the singular value decomposition will be used to calculate intensity profiles for the different HER2-categories and demonstrate the principal modes of variations.

2.8 Using singular value decomposition and shape models for classification

From a feature one could, for a set of known examples, calculate a mean value over this feature. Consider an array of coordinates describing the shape of an object as the feature.

Using singular value decomposition for the set of known examples would then result in the principal directions of variation combined with its singular values. This, together with the mean shape, describes a *shape model* [5]. Suppose that you have the shape model of a given object, one can then approximate a new unknown example (of the same object/category) via the equation

$$\mathbf{S_{new}} \approx \bar{\mathbf{S}} + \mathbf{b_t} * \mathbf{P_t} \tag{6}$$

where $\mathbf{S}$ reflects the shape whereas $\mathbf{t}$ reflects the number of modes of variation that is being used. $\mathbf{P}$ will reflect the principal directions of variation much like V or U in equation 5 but only to a degree such that significant part of the variation (something in the magnitude of 95-99 %) is explained. The $\mathbf{b}$-vector could be calculated by minimising the difference in 6 and reflects how much of the principal variations is needed to approximate the new shape. A common practice is to restrict $\mathbf{b_t}$ to $-k\sqrt{\lambda_t} < \mathbf{b_t} < k\sqrt{\lambda_t}$ where λ in this case corresponds to the singular values in Σ in 5 and k is a constant. To use this equation for classification the idea is that you would approximate each unknown shape with the shape model for each class and then find the best approximation. The best approximation decides the class of the object. This requires that the elements of the feature in S correspond to each other in some way. More specifically meaning, given our example with the set of coordinates, to be able to use shape approximation the coordinates for each shape must clearly correspond to each other. The first coordinate must reflect the same coordinate in each shape.

2.9 Previous work

Image analysis algorithms within digital pathology and breast cancer detection have been developed during the last few years. SlidePath offers a classification algorithm called Tissue IA (TIA) [11] that in general has worked quite well in the past. This algorithm focus on images stained with traditional immunohistochemistry methods. The algorithm isolates the cell-membrane using edge-detection algorithms and uses a colour profile to classify immunopositive tissue. In this way the algorithm can derive a number of metrics related to the ASCO-guidelines such as membrane staining absorbance and percent membrane positive pixels and from these metrics decides the HER2-score.

HER2-CONNECT by Visiopharm [7] is also focusing on traditional IHC stained images. this algorithm also uses a colour profile to extract immunopositive tissue but focuses on the metric *connectivity* in order to asses the HER2-score of the tissue. The connectivity describes the continuous size distribution of immunopositive expression at the cellular membranes and this, means Visiopharm, is highly transferable to the different HER2-scores. This means that the more connected the overall membrane expression is the more probable is a higher score. The connectivity is defined as the area under the curve (AUC) where Y and X are plotted against each other, X and Y are defined by equation 7 and 8. A in these equations would refer to an area of membrane pixels which are stained and demonstrates immunopositive expression.

$$X = \frac{A_{dyn} - A_{min}}{A_{max} - A_{min}} \tag{7}$$

$$Y = \frac{\sum_x A}{\sum A} \tag{8}$$

X would be a ratio between 0 and 1 where A_{dyn} is a dynamical threshold that varies from A_{min} and A_{max} where A_{min} defines membrane stained areas that are too small to be included in the connectivity calculations and where A_{max} defines large membrane segments. Y would also be a ratio between 0 and 1 where the sum in the numerator would refer to all areas above the dynamic threshold A_{dyn} and the sum in the denominator would refer to the total membrane segment area (meaning all membrane segments). The connectivity will depend heavily on the chosen variable A_{dyn} and will be a value also ranging from 0 - 1. The A_{dyn} variable could be calculated using a training set and using purity measures such as Gini impurity index to find the optimal threshold.

Whole slide imaging is not an area that has managed to avoid neural networks and there have been numerous networks developed for HER2-classification in recent years. An example is the deep learning framework her2net developed by Monjoy Saha and Chandan

Chakraborty [23] which reached a very high classification accuracy. This is just one of many neural networks that have been produced in the area but her2net is probably one of the most prominent ones. This a quite complex neural network that starts by cropping out 2048x2048 images from a WSI. Subsequently the cropped out images were further divided into 251x251 patches on which the network was trained on. The patches are fed through a network structure that is explained in short wording below:

1. Convolutional layers: These layers are responsible for learning filters that enhances structural parts of an input images, edges for example.
2. Pooling layers: These layers down-sample convolutional layer output. This is mainly done to keep the number of parameters to learn within reasonable limits.
3. Deconvolutional layers: Much of what it sounds like, deconvolutional layers restores dimensionality to feature maps. This was mainly done in her2net to save the shape of the input image.
4. Fully connected layer is a neural network in which all nodes in one layer are connected to all nodes in the next layer.
5. Dropout layer: This a layer that with a set probability sets randomly selected neurons to zero meaning that some learned features are forgetting. This is done to prevent overfitting.
6. Softmax layer: This is a layer that converts the classificational output to probabilities for each category.
7. LSTM layer: In between the convolutional part and the deconvolutional part the creators of her2net has inserted a Long-short-term memory (LSTM) unit. This unit can be used to insert a long and short term memory in the network. This is a unit often used in language processing for understanding context for a word where the short term could be described to model the sentence whereas the long term the chapter. In this sense, the LSTM-unit was used to remember pixel details from the previous frames. In that way the cropped out patches could be connected to each other and could be held in relation to one another. The LSTM works in simple wording that it takes the previous output from the unit, the current input and the previous hidden state (memory) and combines this to a new output. Each of these inputs are passed through gate functions that makes it possible to either forget or remember them. Saha and Chakraborty also included a trapezoidal LSTM connection to make the correlation between frames dependent on diagonal neighbours.

3 Method

The method will be presented in the following way:

- 3.1 describes the usage of the resource tool QuPath that has been used throughout the project.
- 3.2 describes the data used for the project.
- 3.3 describes how the images were annotated.
- 3.4 describes how the parameters for cell-detection in QuPath were set.
- 3.5 describes the cell-detection and segmentation used.
- 3.6 describes the features used for the classifiers.
- 3.7 describes how the the two different classification models incorporated the features described in 3.6.
- 3.8 describes how the classifiers predicted tissue samples from more granular classification.
- 3.9 describes an overview of the different classifying models used.
- 3.10 describes how the results were visually extracted.

3.1 QuPath

QuPath is an open source image viewer for bioimage analysis and pathology which is being actively developed by Pete Bankhead at Edinburgh University. In this project we have made use of some of QuPaths built in algorithms. One of those is the automatic cell detection algorithm which makes use of traditional edge detection algorithms– Laplacian gradients, Gaussian blurring and watershedding operations– to detect cells. These type of operations have not been further described in the report as they are only used indirectly in our algorithms. The documentation of QuPath could be found on github [22].

3.2 Data

The original data set consisted of 47 tissue samples, in two different modalities: brightfield and UCNP. Three images were removed because of artefacts that made the samples unusable. The reasons for removal were documented preparation issues where the tissue had fallen of the slide or been badly labelled. There was one sample where the reason for removal was that only a very small part of the tissue was usable. The images were obtained with an objective with 20X as the maximal magnification.

Based on the UCNP images, three pathologists classified each tissue into one of the four HER2-categories in the ASCO-guidelines for IHC-classification. The score from the pathologist with the most experience, pathologist 1, of UCNP-images was chosen as the ground truth. The images were divided into a test set and a training set. The test set was chosen at random and consisted of 11 images. All other images were used as training and validation images. The distribution of the tissues' scores, according to pathologist 1, is displayed in Table 1.

Table 1: Distribution of scores in test and training set

Class	Number of tissues in training set	Number of tissues in test set.
HER2-0	9	3
HER2-1+	8	3
HER2-2+	8	4
HER2-3+	8	1

Note that the data-division lacks a validation set. The developed models were only trained on the CPA:s of the training set and then validated against the tissue of the training set for optimising performance. This means the validation set was instead the tissue of the training set. Validation set references in the report refers to the tissue samples of the training set.

3.3 Annotations

Annotations were made following two different procedures: Annotating regions in tissue and annotating whole tissue. The annotations were made using the annotation tools in QuPath. All annotations were conducted on brightfield images of the same tissue sample as the UCNP-images. In that way, the annotations were aligned between the two sets of images.

3.3.1 Annotations of regions of interest in tissue

The images were usually divided into multiple regions of interest (ROI), meaning that an entire tissue sample seldom was regarded as a region of interest. The criteria for a ROI was the following:

1. The tissue was regarded as intact (no cuts or man made holes).
2. The tissue was not of the type fatty tissue.
3. The tissue was cell dense.
4. No presence of artefacts such as bubbles, folds or out of focus areas.

These criteria made the division into ROI:s quite coarse. Since no pathologist was used to annotate the images the strive was to include as much of the tissue as possible. There were also exceptions from the criteria above such as a cell dense area within fatty tissue. In a few cases the entire tissues fulfilled the criteria above. Typically the annotations were made by first assessing the tissue overall to exclude the possibility that the entire tissue was good enough to be deemed as a ROI and to find cell dense areas. Then cell dense areas were subsequently investigated in 200 μm zoom and annotated if they fulfilled the criteria above. In some cases small artefacts or holes were included in the ROI, for example if the ROI was very big a small cut or hole would not impact the ROI significantly. The CPA:s were also annotated but labelled as 'CPA-HER2-X+', where X is the class:(3+, 2+,1+ or 0). Typically each annotation was made by making a polygon around the area of interest. After the annotations were made the built in cell-detection of QuPath was used to extract cell coordinates within the ROI:s. A typical annotation is shown below:

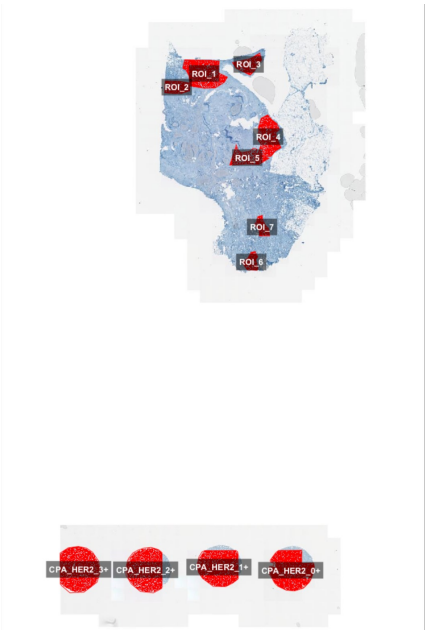


Figure 5: An example of an annotation of a tissue, using ROI-annotations

3.3.2 Annotations of whole tissue

To investigate the impact of our chosen regions and to examine how the models created could be used in clinic, annotations of the whole tissue were made. If pathologists would have to find regions according to guidelines, they could also examine the regions to classify the image, and the total time per image would not decrease significantly. If the models perform as accurately as pathologist the main advantage would then only be that the models do not get fatigued. To gain time, the regions that should be investigated could be found automatically. If only a small part of samples have artefacts, these areas could be so small that the whole tissue could be investigated at once, and the artefacts negligible. If the models still would perform as accurate as on annotated regions, a simple method to find the tissue in brightfield images could be created to automatically find the region of interest.

While conducting the annotation for a complete tissue sample the whole tissue sample was annotated. One restriction that still remained on the tissues were tiles that clearly were out of focus. Since there exists algorithms which can find out of focus tiles with a high accuracy [16], omitting these tiles was possible. Figure 6 displays how a typical tissue was annotated using this method.

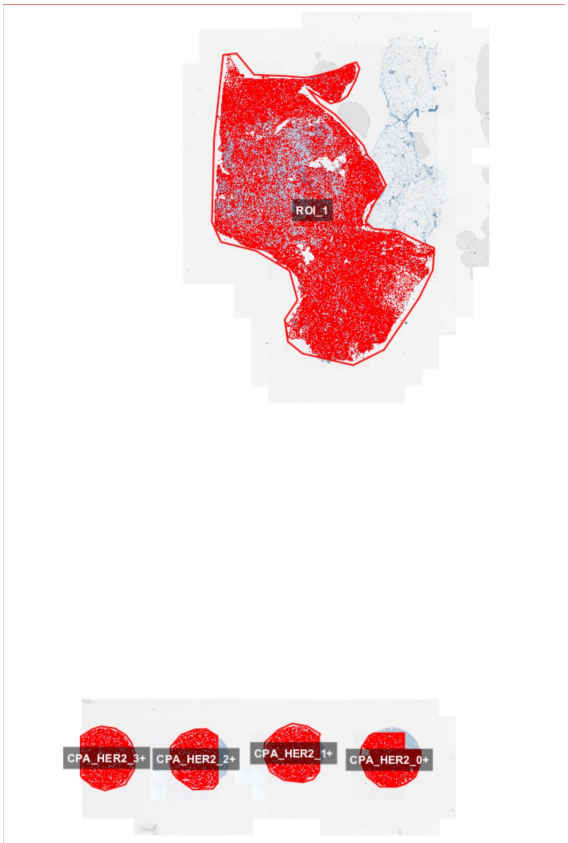


Figure 6: An example of an annotations on the whole tissue sample

3.4 Parameters QuPath cell-detection

The parameters to find the number of cells in QuPath were set according to Table 2

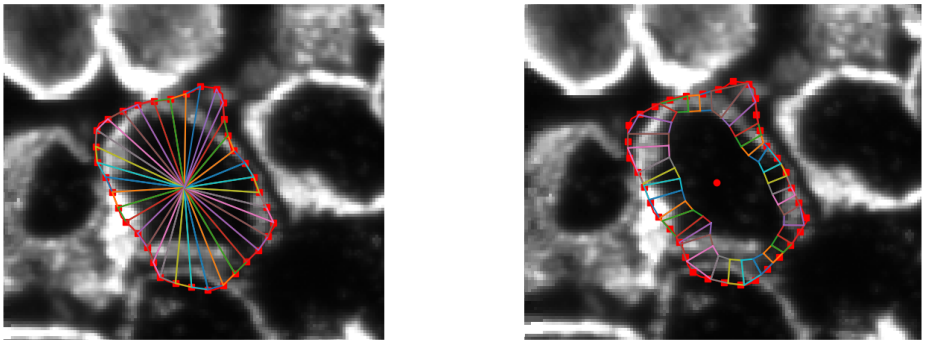
Table 2: Parameters in QuPath

Parameter	Value	Unit
Detection image	Hematoxylin OD	
Requested pixel size	0.5	µm
Background radius	8	µm
Median filter radius	0	µm
Sigma	1.5	µm
Minimum area	10	μm^2
Maximum area	400	μm^2
Threshold	0.1	
Max background intensity	2	
split by shape	box checked	
cell expansion	5	µm
include cell nucleus	box checked	
Smooth boundaries	box checked	
Make measurements	box checked	

3.5 Segmentation of cells

To extract features that are based on the ASCO-guidelines, a segmentation of the cell membrane was desired. With the cell detection algorithm in QuPath, two sets of coordinates were extracted for each cell. One for the nucleus, one for the membrane. It is expected that the protein expression would be located mainly on the outside of the membrane. Therefore, it would be reasonable to investigate the pixels a fixed distance from the membrane. However, the cell detection from QuPath appeared to be slightly inaccurate when transferring the cell-detection from brightfield to UCNP(see appendix). The membrane detection is according to QuPath documentation based on the nucleus detection and is then allowed to grow outwards until another membrane is encountered or until a threshold distance is reached. By empirical investigations the membrane detection was found to surround the membrane rather than exactly correspond to it. To create the segmentation, 36 points were evenly distributed along the membrane derived from QuPath. These 36 points around the membrane of a cell were then used to segment the cell into 36 segments. Two points on the membrane and the center of the nucleus were used to form one segment, creating a triangular shape, see Figure 7a . This method was used to try to include as much of the membrane as possible for each cell uniquely, without covering the same area twice. Meaning that the goal for each segmentation was that it corresponded to one and only one cell. In this way one could investigate features that are derived from the attributes from cellular membranes such as, for example, intrasegmential intensity variance and use these to correlate to ASCO-guidelines. Other cell-segmentations were investigated such as finding the orthogonal intersect between the line between two points, of evenly distributed points, along the nucleus and the membrane. Think of it as drawing an orthogonal line from the nucleus segmentation to the membrane and finding the intersection point. One would then use pair of two intersection points, with two points of nucleus points as a four point polygon segment. This however provided to be quite difficult to use since cells have very wide variations of shapes which creates strange intersections with the membrane. Consider a cell in mitosis for example. This segmentation was also more complex than the one previously described, required more computations and in general performed worse in segmenting the membrane correctly which led to the choice of segmentation method.

The segmentation methods that were investigated are presented below. The one that was used is the left one of the two figures. The right one is the method that was discarded due to complexity and the lack of robustness to irregular shapes.



(a) Segmentation method 1 (b) Segmentation method 2

Figure 7: Segmentation methods

3.6 Feature extraction

Features were extracted on two levels. Either based on entire regions of interest or, based on the cells within a region.

3.6.1 Features on regional level

From the region mainly one feature was investigated and called the *normalised intensity* which was the intensity within the region normalised with the cell count within that region. The summed intensity of each pellet on the region was computed by multiplying a binary mask of the annotation with the image itself and then summing all the intensities. The binary mask was built up as a squared matrix with the height and length of the annotation, its pixels were 1 if the pixels position was within the annotation and 0 if the pixels position was outside the annotation. The summed intensity was normalised with the number of detected cells in the same region. A normalisation was necessary since the UCNP particles binds to cell membrane, the intensity is therefore highly correlated to the number of cells. The theory was that each class would be within a specific range with an intensity per cell that corresponded to the amount of HER2 proteins attached to the membrane. HER2-3+ would have the highest intensity per cell, and then a decreasing intensity per cell for the rest of the classes.

3.6.2 Features on cellular level

From the cell segmentation shown above, Figure 7, a number of different features were examined. The purpose of the features was to capture the coverage around the cell as theory has shown that HER2-positive cells has a coverage of the HER2 expression along a majority of the membrane. To capture this mainly two features were examined:

1. Mean of the top x % of the segment
2. Variance of the sequential difference

The mean of the top x % of the segment simply extracted the average intensity of the top pixels in the segment. The idea was that a HER2-3+ would have high values in all the segments and express stable high values on average, a HER2-2+ would have more variance around the cell and occasionally express high values and occasionally express low values in the segments, a HER2-1+ would follow roughly the same pattern as the 2+ but with lower values whereas the HER2-0 would on average express low stable values that varies with noise. This feature was used as the top values should correlate to the membrane pixels that should exist somewhere in the segment. The mean value was considered to be too susceptible

to non membrane pixels. Using the mean of the top % pixel values instead was considered to be robust to this feature, as long as there were any membrane pixels in the segment. Using only a max value would enable the possibility for spurious pixels to indicate high HER2-expression. For each cell one would extract an array with 36 values corresponding to the value for this feature in the segment.

The variance of the sequential difference was a feature derived by calculating the absolute difference between the mean intensity of one segment and the mean intensity of the next segment and then calculating the variance over all the differences. This resulted in one variance per cell. The objective of this feature was to find the change in coverage around cells. According to the ASCO-guidelines, a tissue that is classified as a HER2-0 has no or faint incomplete membrane staining, therefore the variance of the sequential difference between segments should be low for this category. A standard HER2-1+ cell, has a faint incomplete membrane staining and the difference between no staining and faint is expected to be large. A typical HER2-2+ cell have complete staining, and varies between moderate and weak, an intensity difference that is expected to be larger than the difference between segments for HER2-1+. Lastly, a HER2-3+ cell usually have a complete intensity that is intense around the whole cell. If it is intense around the whole cell, the difference between each segment is expected to be low, and the variance of the sequential difference would therefore be low.

Probability density plots were created for features from CPA:s in the training set to examine there individual ability to differentiate between different classes. These gave a good indication of the features strength and could be used to explain the final predictions using this algorithm. A 3D-plot was also constructed combing all the features normalised to investigate separation.

3.7 Classifiers

3.7.1 Training, validation and prediction

All classifiers have in common that they were trained on the CPA:s in the training set, since ground truth data on specific parts of a tissue not was available nor is a label for a tissue sample applicable to the entire sample. The classifiers were validated on tissue, as described in section 3.8, in the training set and subsequently tested on both CPA:s and tissue in the test set. The classifiers were, after test set prediction, retrained on all of the images to investigate optimal performance. This was done using both ROI-annotation and complete tissue annotation (CPA-prediction was the same for both annotations). The classifiers performance, on both CPA:s and tissue, were evaluated via confusion matrices.

After completing all steps above a final k-fold cross validation over the dataset was done dividing the dataset, randomly, into 4 folds with 11 samples in each. This led to four set of validation scores, one for each fold where the models were trained on all the samples except the ones in the current fold. This was done for all the models using ROI-annotations. The accuracy for all the folds as well as the mean accuracy was reported.

3.7.2 Gaussian naive Bayesian classifier using normalised intensity

Statistical properties, from CPA:s, necessary to conduct a Bayesian classifier were extracted for each class based on the normalised intensity. The classification algorithm then proceeded to use the learned statistical properties of each category regarding the normalised intensity to classify new, unseen samples by making use of the Bayesian classifying algorithm described in equation 4. Ground truth data exists for each pellet (HER2-0, HER2-1+, HER2-2+ and HER2-3+) in the CPA. Therefore, each pellet could be predicted using the Bayesian classifying algorithm.

Evaluating the models performance on real tissue could not be created in a similar way since no labels for each annotated tissue-part was available. Instead, each annotated region from a tissue was divided into squares of size 512 x 512 pixels (167 x 167 µm). All an-

notated regions were divided into several patches, and each patch could then be classified with the Bayesian model. The patches were processed as long as they intersected with the annotation and contained more than 50 cells. The number of cells needed for processing were empirically decided to approximately a quarter of the number of cells in a patch in the middle of an annotation. The classified patches could then be combined according to the methodology described in section 3.8, to predict the class of a whole tissue sample.

3.7.3 Multidimensional Bayesian classification

To increase the performance of the Bayesian model with normalised intensity, more features were added and combined with the normalised intensity features to create a multidimensional Bayesian classifier. The added features used were the ones described in section 3.6.2. In contrast to the Bayesian classifier using normalised intensity, these features were more detailed and granular. Using only normalised intensity, only one number per region was found and could be used as an input to the Bayesian classifier. The mean of the top 5% feature was calculated for each cell in a region, all cells within a region therefore had 36 values. To reform this 36-dimensional feature per cell to instead one value per region the following was done:

1. The average value over the 36 values was extracted. Creating an average top value per cell.
2. The average of this average value was extracted over all cells. Creating an average top value for the region.

This value could then be used as an input to the Bayesian model. For the feature "variance of sequential difference", one value for each cell existed and an average over all the cells was found to be used as input to the Bayesian model.

3.7.4 Classifier using intensity profiles

For each segment in a cell the mean of the top 5 % pixel values, as described in section 3.6, were used to create an intensity profile for a cell. Over all of the cells in the CPA:s on the training images a SVD based on this metric was made using equation 5. This created a mean intensity profile for each HER2 category along with principal modes of variation and corresponding principal vectors, an intensity profile model. This intensity profile is analogous to the shape in equation 6 whereas the intensity profile model is analogous to the shape model. To remove the spatial dependence the intensity values along the segments were sorted. An unsorted example intensity profile is shown below (the values are simulated and not real).

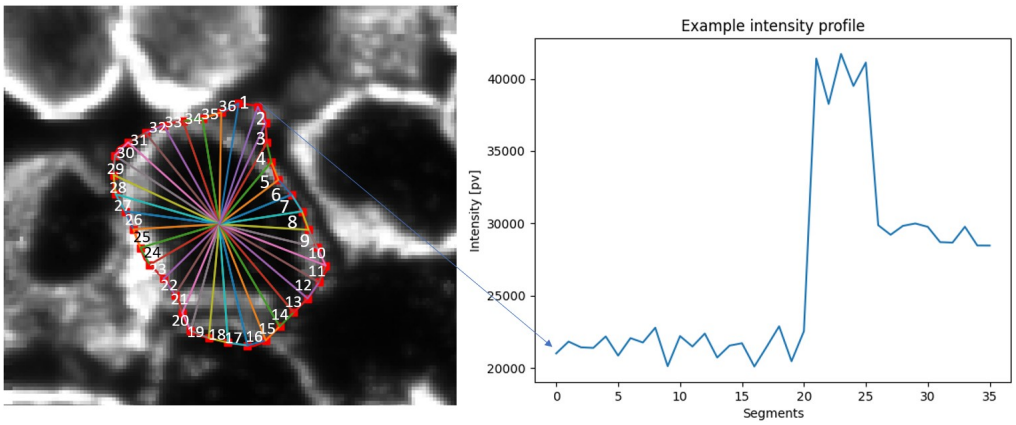


Figure 8: Example of unsorted intensity profile

An intensity profile classifier would classify individual cells, where ideally the cells of the different HER2-categories expresses different intensity profiles. A 3+ cell would theoretically express an intensity profile that is more flat (high values over all coordinates), with

high values and a 0+ cell would express a more noisy intensity profile with lower values. Given the intensity profile of a new cell an approximation for the intensity profile model for each class was made following equation 6 and the best approximation decided the class of that cell. The best approximation was decided using mean squared error (MSE) over the segments. Based on all cell classifications in a sample, a whole tissue could be classified according to section 3.8.

The top percentage to be used was set to 5 % since this optimised the performance over the CPA:s in the training set. The percentages 5, 10, 15, 20 and 25 were investigated over all of the CPA:s. For each CPA in all samples in the training set, 10 % of the cells (227151) were used to create an intensity profile model for all of the different categories for the different percentages. Then the calculated intensity profile models were used for predicting the same cells that were used for deriving the intensity profile models in the first place. The percentage that minimised the number of misclassifications was chosen and was deducted to 5 % (this percentage was applied on the previous described multi dimensional Gaussian classifier as well). The intensity profile models were then built up using all the CPA:s in the training set mounting up to approximately 528000 (33 x 16000) cells per HER2-class.

As mentioned a common practice is to restrict $\mathbf{b_t}$ to $-k\sqrt{\lambda_t} < \mathbf{b_t} < k\sqrt{\lambda_t}$ where λ corresponds to the singular values in Σ from equation 5. This restriction k was set to 2.1 as this minimised the error over the CPA:s in the training set. Restriction coefficients from 0-4 were investigated and gradually increasing the granularity. Meaning, that first increments of 0.5 were used and once the minima had been located in between some values these were used as boundaries and the increment size decreased. This continued until increments of 0.01 and the boundaries of 2.0 and 2.2 were used.

An total intensity profile model for all the cells in the CPA:s (regardless of HER2-category) was also made to illustrate the largest modes of variation. The idea was that they would correspond to the ASCO-guidelines.

3.8 Prediction methodology

After prediction on patch or cell level, a model to predict the whole tissue was created. This model combined all predictions on cell level for the intensity profile classifier and on patch level for the Bayesian classifiers.

3.8.1 Prediction of tissue from patches

Based on the classification of patches from Bayesian classifier a whole tissue sample could be predicted by combining the patches in two different ways. The patches were either counted, and the absolute quantity of patches within each class in a tissue was found, or the percentage of each class per tissue was found. Counting the number of patches could be highly dependent on the area of all annotated regions within a tissue. On the other hand, it could give an absolute value for the quantity of patches needed within a specific class to predict the whole tissue sample to that class. By finding the percentage of patches within a region, it will be normalised with the area of all regions in each sample. Therefore, the size of the annotated area should not impact the classification. However, the distribution of HER2 proteins in a tissue can vary, and a class can be determined according to only a small subsection of the whole tissue. Therefore normalising with all regions could disturb the actual results. Because of this, both models were investigated.

3.8.2 Prediction of tissue from cells

Based on the cell classification a score for the entire sample was deducted. This deduction was done following two different metrics, either using the ratios between the cell classifications in the tissue sample as the decision metric or by directly using the absolute quantity of cells within the different classes as the decision metric. The former was further divided into two branches: Full sample ratio and the per ROI mean ratio. The difference between the two methods is that the latter is focusing on the ratio within the ROI:s. This alternative was

created in order to not overly average out high-class regions over the sample. To summarise, the following methods were used to decide sample classification:

1. Using the ratio of cells in the different categories.
2. Using the mean ROI-ratio of cells in the different categories.
3. Using the absolute quantity of cells in the different categories.

3.8.3 Prediction of tissue

The number of predicted patches or cells were used to predict the whole tissue. Threshold values were found to decide how many predictions of each class that was needed to classify the whole tissue as that class. The model first investigated the absolute quantity or ratio of HER2-3+ prediction, if the calculated value was above the threshold value, the tissue was predicted as a HER2 3+. If the quantity or ratio for HER2-3+ was below the threshold value, the model investigated if it could be classified as a HER2-2+, using both the predicted HER2-3+ and HER2-2+ values. The same applied for HER2-1+-classification. If the ratio or quantity of HER2-3+ predictions, HER2-2+ predictions and HER2-1+ predictions surpassed the HER2-1+ threshold the sample was classified as HER2-1+. In this way, the whole tissue could be classified. In total 3 threshold values were created, one for HER2-3+, one for HER2-2+, and one for HER2-1+, if none of the thresholds were surpassed, the sample was classified as HER2-0. In this way, the prediction of tissue from cells or patches resembled how a pathologist would examine a sample according to the ASCO-guidelines, classifying one sample at a time.

3.8.4 Finding the threshold values

The optimal threshold values were found by finding the threshold values that gave the best accuracy over the training set. While calculating the ratio thresholds, the algorithm iterated over 100 threshold values for each threshold, and saved the thresholds that derived the best accuracy on the training set. While finding the thresholds for absolute quantity a suitable range and step size was chosen for each model individually. The range for each threshold value was first linked to the maximal number of patches or cells found in the training set. Finding an indication what the threshold values were, the range could be adjusted to decrease the step size and in that way find more accurate threshold values. The final ranges and step sizes for the models can be found in the Appendix C. Thereafter these threshold values could be used on the test set. After test set prediction, all images where considered to be available as training material and new thresholds where deducted over the entire set of images to investigate threshold stability. When predicting on CPA:s the threshold values (if for the classifier applicable) optimised for CPA-prediction were used.

3.9 Overview of final classifying models

Table 3: Abbreviation explanation

Model abbreviation	Model
NGM 1	Naive Gaussian model using absolute quantity as threshold type
NGM 2	Naive Gaussian model using ratio as threshold type
MDM 1	Multi dimensional Gaussian model using absolute quantity as threshold type
MDM 2	Multi dimensional Gaussian model using ratio as threshold type
IPC 1	Intensity profile classifier using whole sample cell ratio as threshold type
IPC 2	Intensity profile classifier using mean ROI-cell-ratio as threshold type
IPC 3	Intensity profile classifier using absolute quantity as threshold type

3.10 Visual presentation

Predictions from the classifiers were also extracted in a QuPath compatible format called geojson that could be overlaid on UCNP-images to provide hypothetical visual decision support for a pathologist. The visual results were also reported alongside the qualitative results.

4 Results

In this section the results will be presented, both visually and statistically displayed. For the confusion matrices the accuracy will be given in parentheses underneath each matrix.

4.1 Visual overlook

In this section the visual overlook for the ROI-annotation will be presented. An example of visual results for the complete tissue annotation is found in Appendix A.2.

4.1.1 Sample 1 at different zoom levels

Below, Figure 10-12 a visual overlook of how the multidimensional Bayesian classifier and the intensity profile classifier classifies the patches and cells is presented on different zoom levels. The information from the intensity profile classifier will be shown to the right in the subfigures and the information from the Bayesian classifier to the left. The cells and patches are labelled according to Table 4.

Table 4: Class and colour scheme for images

Class	Colour
HER2-3+	Red
HER2-2+	Orange
HER2-1+	Green
HER2-0	Blue

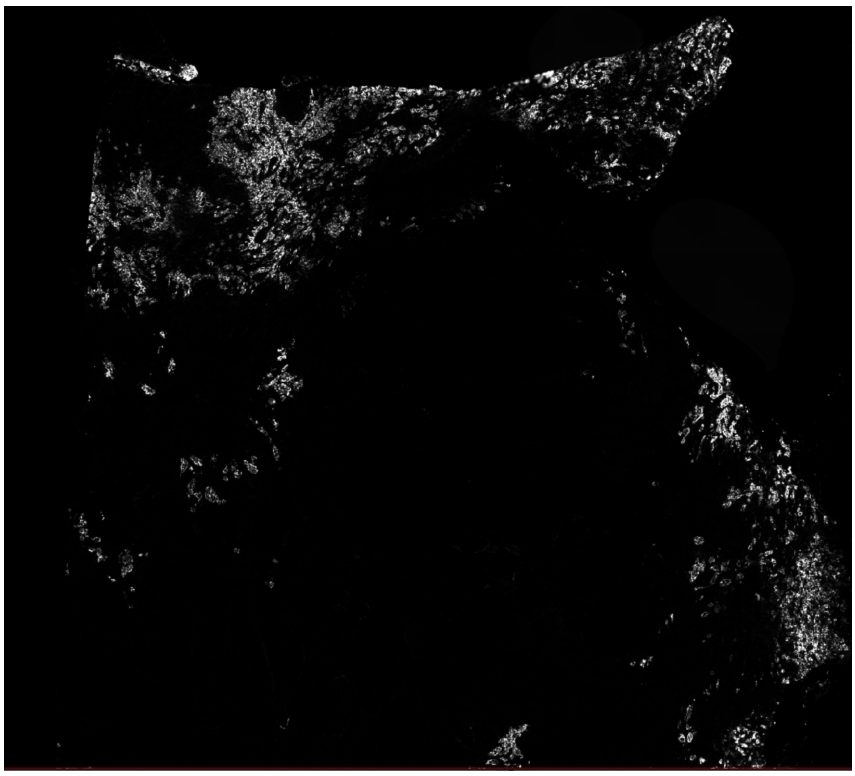
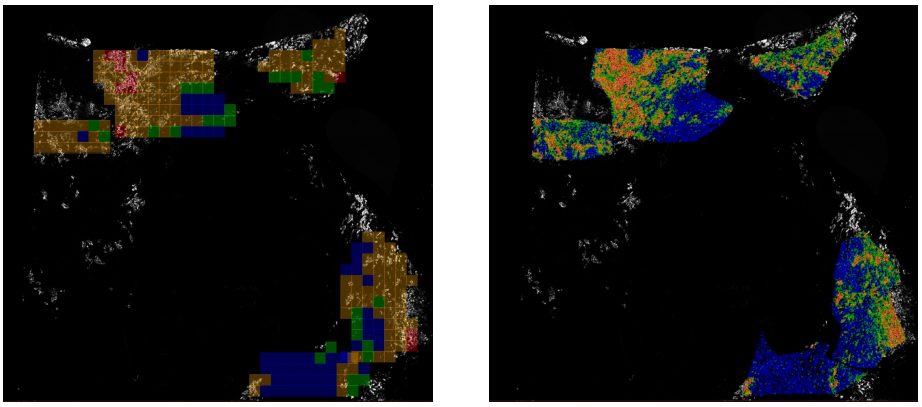
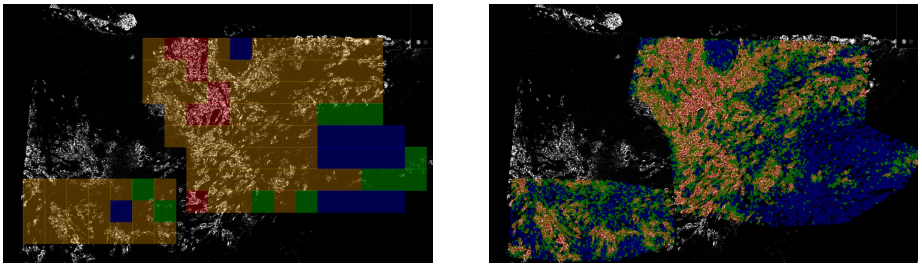


Figure 9: A UCNP-image without any classification information



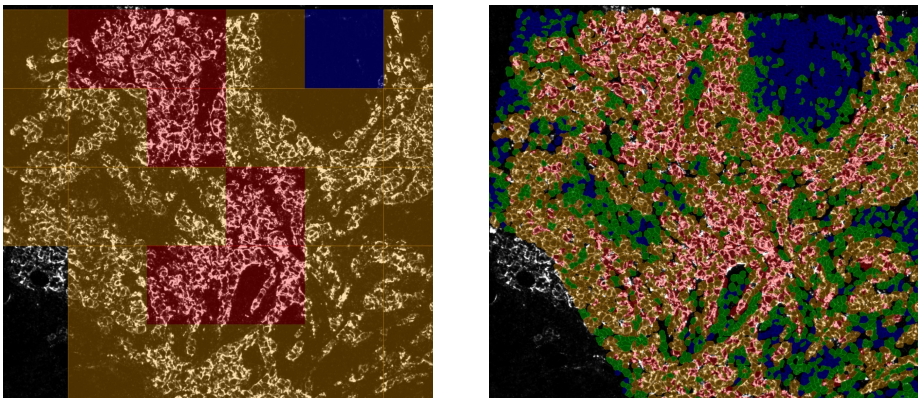
(a) Patch classification overlay (b) Cell classification overlay

Figure 10: Figure 9 with classification overlays



(a) Patch classification overlay zoom 1 (b) Cell classification overlay zoom 1

Figure 11: Figure 10 in zoom level 1

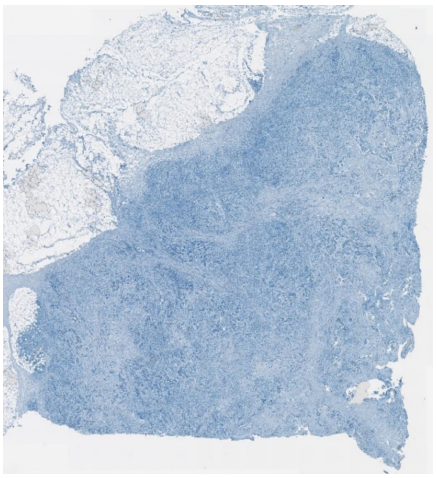


(a) Patch classification overlay zoom 2 (b) Cell classification overlay zoom 2

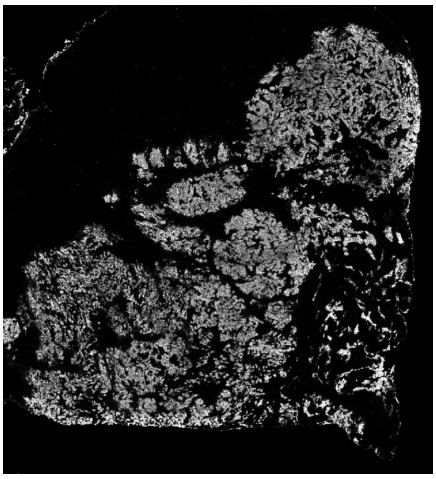
Figure 12: Figure 10 in zoom level 2

4.1.2 Visual results per class

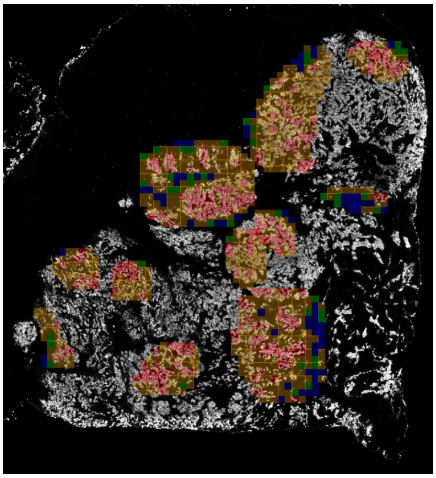
Figure 13-16 displays four tissue samples that both models have predicted correctly. To the top left the tissue sample in brightfield is presented, followed by the tissue sample in UCNP without any overlays to the top right. To the bottom left the patch classification from the Gaussian classifier is overlaid and to the bottom right the cell classification from the intensity profile classifier is overlaid.



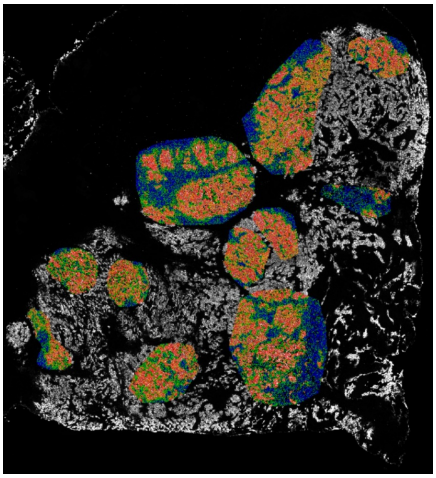
(a) 3+ image in brightfield



(b) 3+ image without overlay

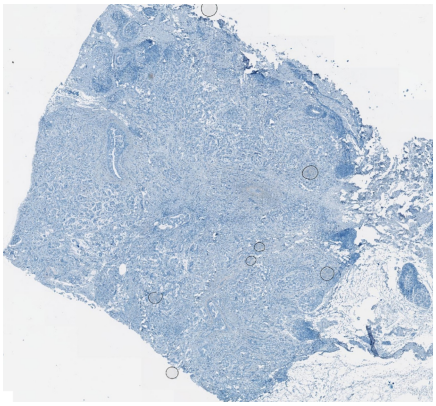


(c) Patch classification overlay

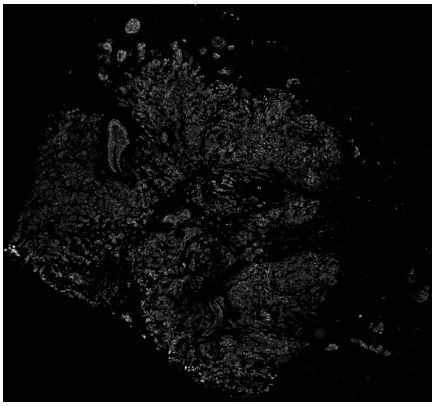


(d) Cell classification overlay

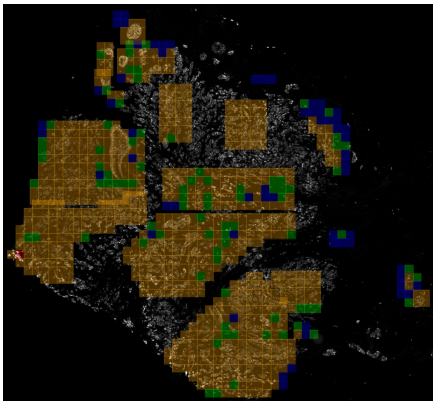
Figure 13: 3+ sample with different classification overlays



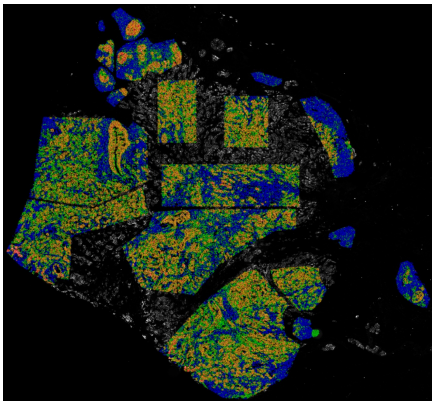
(a) 2+ image in brightfield



(b) 2+ image without overlay

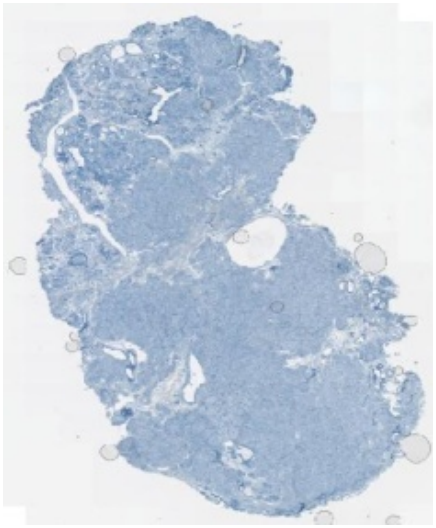


(c) Patch classification overlay



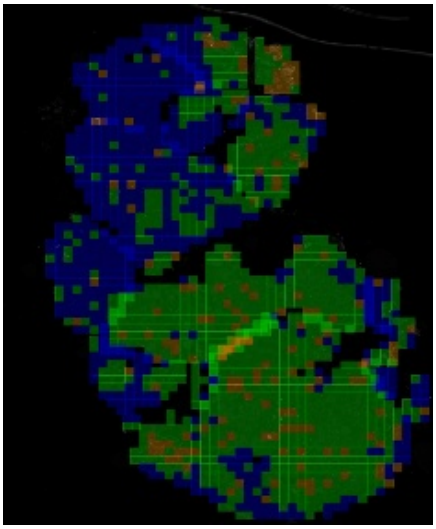
(d) Cell classification overlay

Figure 14: 2+ sample with different classification overlays

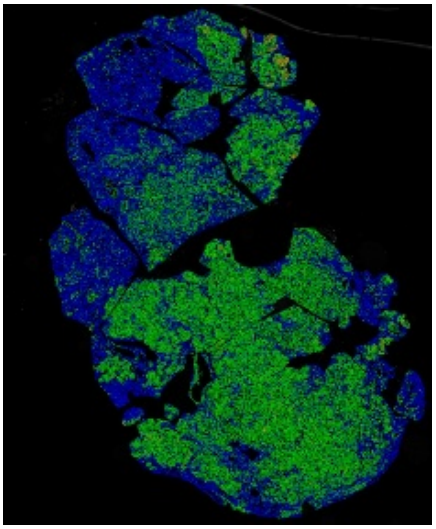


(a) 1+ image in brightfield

(b) 1+ image without overlay

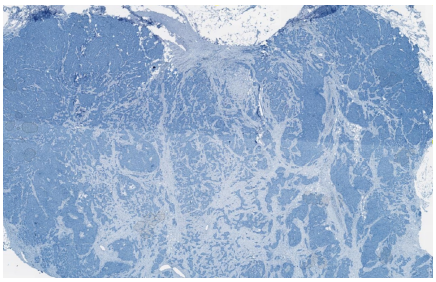


(c) Patch classification overlay



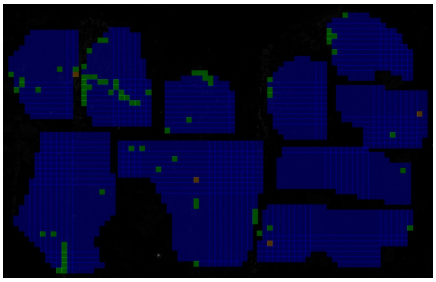
(d) Cell classification overlay

Figure 15: 1+ sample with different classification overlays

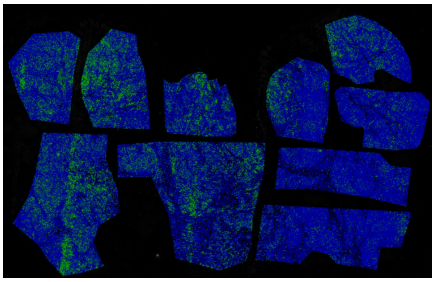


(a) 0 image in brightfield

(b) 0 image without overlay



(c) Patch classification overlay

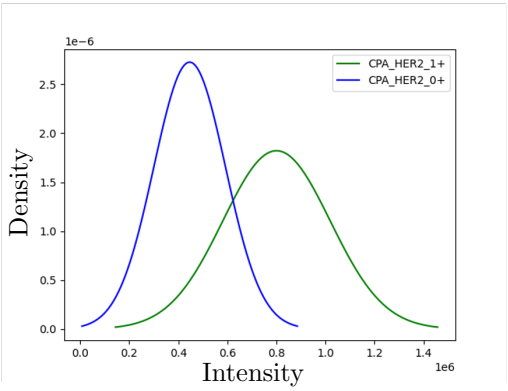


(d) Cell classification overlay

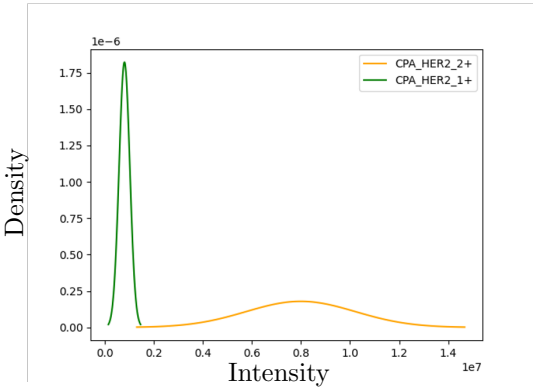
Figure 16: 0 sample with different classification overlays

4.2 Bayesian models

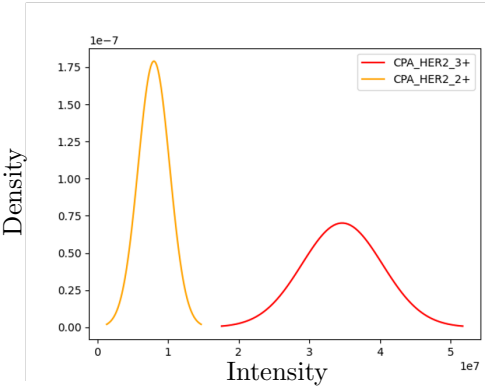
In Figure 17-19 probability density functions for all 4 classes are displayed over all features used for the Bayesian models. The features are; normalised intensity, mean of top values and sequential segment variance. As the images displayes, there are clear differences for all features between HER2-3+ and HER2-2+, as well as HER2-2+ and HER2-1+. However, there is a large overlap for HER2-1+ and HER2-0, for all features.



(a) HER2-0 in blue and HER2-1+ in green

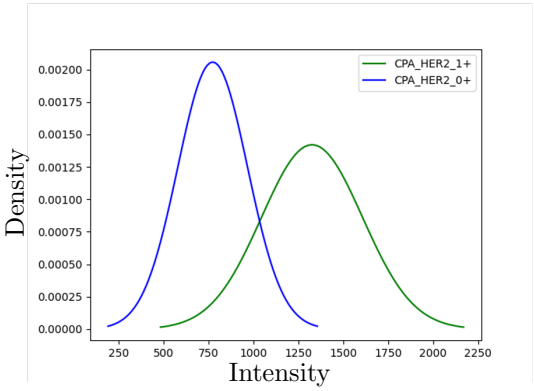


(b) HER2-1+ in green and HER2-2+ in yellow

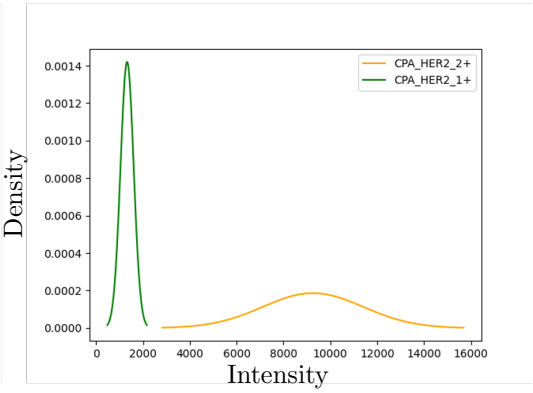


(c) HER2-2+ in yellow and HER2-3+ in red

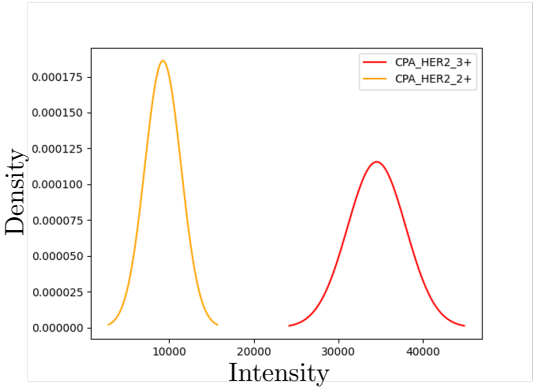
Figure 17: Probability density functions for Normalized intensity for CPA:S in the training set



(a) HER2-0 in blue and HER2-1+ in green

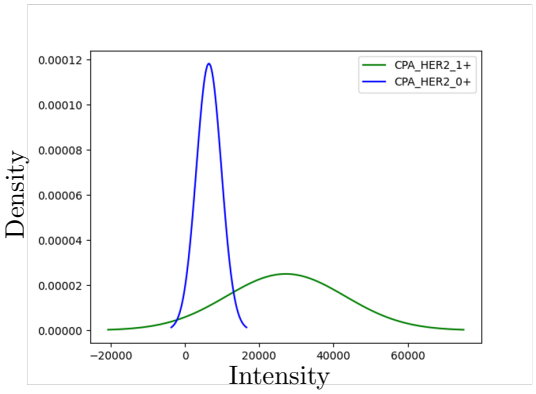


(b) HER2-1+ in green and HER2-2+ in yellow

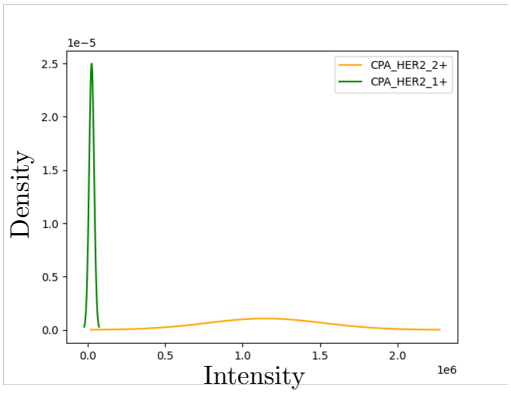


(c) HER2-2+ in yellow and HER2-3+ in red

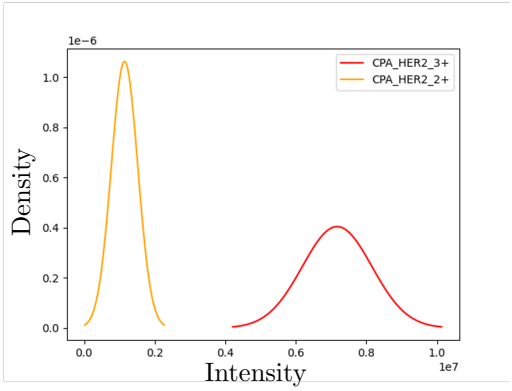
Figure 18: Probability density functions for mean of top values for CPA:S in the training set



(a) HER2-0 in blue and HER2-1+ in green



(b) HER2-1+ in green and HER2-2+ in yellow



(c) HER2-2+ in yellow and HER2-3+ in red

Figure 19: Probability density functions for variance of sequential difference for CPA:S in the training set

In Figure 20 the featuires of the CPA:s have been plotted in a normalised 3 dimensional feature space. In the plot red dots are HER2-3+, yellow are HER2-2+, blue are HER2-1+ and green are HER2-0. The plot displays that the HER2-3+ and HER2-2+ category is separable to the other categories, but HER2-0 and HER2-1+ are very similar, even when combining all features.

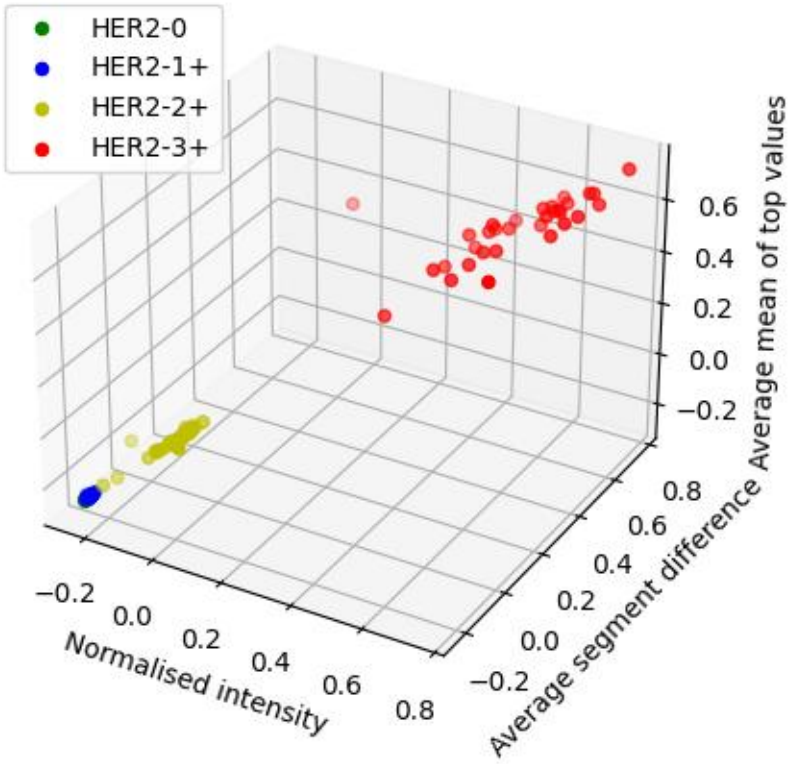


Figure 20: Plot of the CPA:s in the training set in the 3 dimensional feature space. In this feature space HER2-3+ pellets are red, HER2-2+ pellets are yellow, HER2-1+ pellets are blue and HER2-0 pellets are green

4.2.1 Predictions on CPA

Prediction result on the training set, using both Bayesian models, is visualised as a confusion matrices in Figure 21. Figure 22 displays the confusion matrix on the test set. The test set in this instance will consist of 44 CPA-pellets (11 images in the test set with 4 pellets, of each class, each). All confusion matrices have predicted classes with the models designed in the columns and the true classes in the rows. As Figure 21 displays, both models have a very similar distribution across the confusion matrix and comparable accuracy in the training set. Figure 22, displays that both Bayesian models classify the samples in an identical way.

	3+	2+	1+	0
3+	31	1	0	0
2+	0	32	0	0
1+	0	0	28	4
0	0	0	3	29

(a) **Normalised intensity** Bayesian classifier (93.75 %)

	3+	2+	1+	0
3+	32	0	0	0
2+	0	32	0	0
1+	0	0	27	5
0	0	0	1	31

(b) **Multidimensional** Bayesian classification (95.3 %)

Figure 21: Confusion matrices on the training set from prediction of CPA:s using two variants of the Bayesian Classifier

	3+	2+	1+	0
3+	11	0	0	0
2+	0	11	0	0
1+	0	0	10	1
0	0	0	0	11

(a) **Normalised intensity** Bayesian classifier (97.7 %)

	3+	2+	1+	0
3+	11	0	0	0
2+	0	11	0	0
1+	0	0	10	1
0	0	0	0	11

(b) **Multidimensional** Bayesian classification (97.7 %)

Figure 22: Confusion matrices on the test set from prediction of CPA:s using two variants of the Bayesian Classifier

4.2.2 Prediction on tissue using ROI-annotaions

The threshold values that optimised the accuracy for both the training set and for all images, using the Bayesian classifier with normalised intensity, is displayed in Figure 23. In Figure 24, the threshold values for both training set and all images, using the multidimensional Bayesian classifier, are displayed. The values in the column "Training Set" and the values in column "All Images" in Figure 23 and 24 do not vary very much in any of the models except the Bayesian classifier using normalised intensity and ratio as threshold value.

Threshold	Training Set	All Images
3+	8	12
2+	60	60
1+	120	120

(a) Absolute quantity

Threshold	Training Set	All Images
3+	0.01	0.01
2+	0.25	0.25
1+	0.07	0.38

(b) Ratio

Figure 23: Threshold values that optimised confusion matrices for both training set and all images, using Bayesian classifier with **normalised intensity**

Threshold	Training Set	All Images
3+	7	10
2+	170	170
1+	80	90

(a) Absolute quantity

Threshold	Training Set	All Images
3+	0.01	0.01
2+	0.27	0.27
1+	0.15	0.15

(b) Ratio

Figure 24: Threshold values that optimised confusion matrices for both training set and all images, using Bayesian classifier with **multiple features**

In Figure 25 the optimal confusion matrices, using both different threshold types, can be examined, using the normalised intensity as the only feature, with patch size 512 x 512 pixels. In Figure 26 the optimal confusion matrices, using both different threshold types, for a Bayesian model using multiple features is displayed. All models perform approximately with the same accuracy, and the confusion matrices have a similar distribution. The most misclassifications occur between classification of 0 and 1+, as well as between 3+ and 2+.

	3+	2+	1+	0
3+	7	2	0	0
2+	2	6	0	0
1+	0	0	6	2
0	0	1	3	4

(a) Absolute quantity (69.7%)

	3+	2+	1+	0
3+	7	1	1	0
2+	2	6	0	0
1+	0	0	7	1
0	0	0	4	4

(b) Ratio (72.7%)

Figure 25: Confusion matrix of prediction on training set using both absolute quantity(a) and ratio(b) as threshold values, with Bayesian classifier using only **normalised intensity**.

	3+	2+	1+	0
3+	8	1	0	0
2+	2	6	0	0
1+	0	0	8	0
0	0	0	5	3

(a) Absolute quantity (75.6%)

	3+	2+	1+	0
3+	7	1	1	0
2+	3	5	0	0
1+	0	0	7	1
0	0	0	3	5

(b) Ratio (72.7%)

Figure 26: Confusion matrix of prediction on training set using both absolute quantity(a) and ratio(b) as threshold values, Bayesian Classifier using **multiple** features.

In Figure 27 and Figure 28 the confusion matrices on the test set for the two Bayesian models are displayed using both ratio and number of patches required as thresholds. As the confusion matrices in Figure 27 and 28, both models that used absolute quantity perform significantly better than the models that are using ratio as threshold. It is also clear that the Multidimensional Bayesian Classifier performs better than the Bayesian Classifier that only uses normalised intensity, compare Figure 27 (a) with 28 (a), and Figure 27 (b) with 28.

	3+	2+	1+	0
3+	1	0	0	0
2+	2	2	0	0
1+	0	0	3	0
0	0	0	1	2

(a) Absolute quantity (72.7%)

	3+	2+	1+	0
3+	1	0	0	0
2+	2	2	0	0
1+	0	0	2	1
0	0	0	3	0

(b) Ratio (45.5%)

Figure 27: Confusion matrix of prediction on test set using both absolute quantity(a) and ratio(b) as threshold values, with Bayesian classifier using only **normalised intensity**.

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	3	0
0	0	0	1	2

(a) Absolute quantity (81.8%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	2	1

(b) Ratio (63.6 %)

Figure 28: Confusion matrix of prediction on test set using both absolute quantity(a) and ratio(b) as threshold values, Bayesian Classifier using **multiple** features.

If all images were used to find the optimal thresholds, confusion matrices 29 and 30 could be created. Compared to 27 and 28, the accuracies for all models have improved, and the differences in accuracies between models are not as large as it was in the test set.

	3+	2+	1+	0
3+	8	2	0	0
2+	3	9	0	0
1+	0	0	9	2
0	0	1	4	6

(a) Absolute quantity(72.7%)

	3+	2+	1+	0
3+	8	1	1	0
2+	4	8	0	0
1+	0	0	4	7
0	0	0	1	10

(b) Ratio (68.2%)

Figure 29: Confusion matrix of prediction on all images set using both absolute quantity(a) and ratio(b) as threshold values, with Bayesian classifier using only **normalised intensity**.

	3+	2+	1+	0
3+	9	1	0	0
2+	3	9	0	0
1+	0	0	11	0
0	0	0	6	5

(a) Absolute quantity (77.3%)

	3+	2+	1+	0
3+	8	1	1	0
2+	5	7	0	0
1+	0	0	9	2
0	0	0	5	6

(b) Ratio (68.2%)

Figure 30: Confusion matrix of prediction on test set using both absolute quantity(a) and ratio(b) as threshold values, with Bayesian Classifier using **multiple** features.

4.2.3 Prediction on tissue using complete tissue annotation

Prediction results using images that with complete tissue annotation and predicted with the Bayesian classifier is displayed in this section. Results using all images as a training set for this annotation is found in Appendix B.

In Figure 31 the optimal threshold values are displayed for the two Bayesian models. Compared to the threshold values using ROI-annotations for the training set, see Figure 23 and 24, the threshold values have changed for some of the thresholds.

Threshold	Ratio	Absolute quantity
3+	0.01	6
2+	0.13	230
1+	0.07	180

(a) Threshold values for Bayesian model 512 x 512, only **normalised intensity**

Threshold	Ratio	Absolute quantity
3+	0.01	14
2+	0.18	180
1+	0.12	220

(b) Threshold values for Bayesian model 512 x 512, **multidimensional**

Figure 31: Bayesian classifier thresholds on **complete** annotated images

Figure 32 & 33 displays the confusion matrices for the Bayesian models predictions on the training set, using absolute quantity and ratio as threshold criteria. The confusion matrices and the accuracies using complete annotations are similar to the results for ROI-annotations in the training set for all Bayesian models, compare Figure 32 & 33 with Figure 25 and 26. The accuracies do not vary more than 6 percentage points, and the distribution in the confusion matrices have the same tendencies.

	3+	2+	1+	0
3+	8	0	1	0
2+	3	5	0	0
1+	0	0	7	1
0	0	0	5	3

(a) Absolute quantity (69.7%)

	3+	2+	1+	0
3+	5	3	1	0
2+	1	7	0	0
1+	0	0	8	0
0	0	0	4	4

(b) Ratio (72.7%)

Figure 32: Bayesian classifier using only **normalised intensity** for complete annotated images

	3+	2+	1+	0
3+	8	1	0	0
2+	2	6	0	0
1+	0	1	7	0
0	0	0	5	3

(a) Absolute quantity (72.7%)

	3+	2+	1+	0
3+	7	1	1	0
2+	2	6	0	0
1+	0	0	7	1
0	0	0	3	5

(b) Ratio (75.7%)

Figure 33: Bayesian classifier using only **multidimensional** for complete annotated images

Figure 34 & 35 displays the confusion matrices for the Bayesian models predictions on the test set, using both number of patches and ratio as threshold criteria. Compared to the test set for ROI-annotations the accuracies using complete models are equal to or approximately 10 percentage points better than the accuracies using ROI-annotations. The model still only misclassifies between 1+ and 0, as well as 3+ and 2+.

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	3	0
0	0	0	2	1

(a) Absolute quantity (72.7%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	3	0

(b) Ratio (54.5%)

Figure 34: Bayesian classifier using only **normalised intensity** for complete annotated images

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	3	0
0	0	0	1	2

(a) Absolute quantity (81.8%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	1	2

(b) Ratio (72.7%)

Figure 35: **Multidimensional** Bayesian classifier for complete annotated imaged

4.3 Intensity profile classifier (IPC)

In the table below the different error numbers over classifications of 10 % of the cells in the CPA:s in the training set using an intensity profile classifier derived with the different top x percentages of the membrane segments are presented:

Table 5: Errors over 227151 cells for intensity profiles derived with mean of of the top x % of the membrane segments

Percentage of segment [%]	Error number
5	49333
10	49874
15	50643
20	51415
25	52179

The result below illustrates the mean intensity profile for all the cells in the CPA:s as if they belonged to the same class (no HER2-class separation) and the modes of variation that explains 99 % of the variation. This shows that biggest modes of variation explains the slope profile and the position of the slope. The first mode of variation describes where the slope is positioned, the magnitude of the intensity values. The second mode of variation describes how the slope behaves, the linearity of the slope.

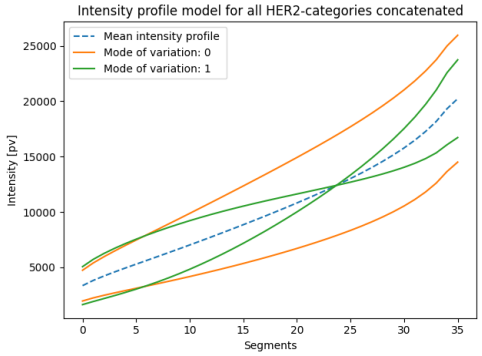
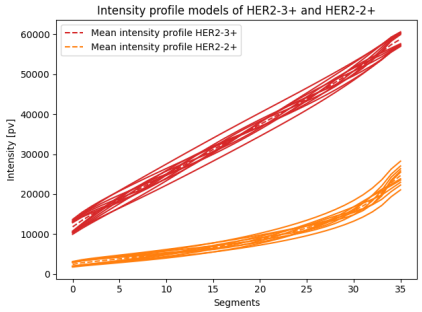
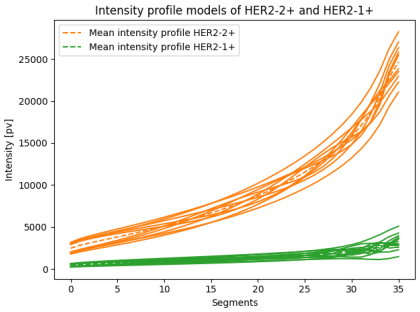


Figure 36: The mean intensity profile for all cells in the CPA:s with the primary modes of variation

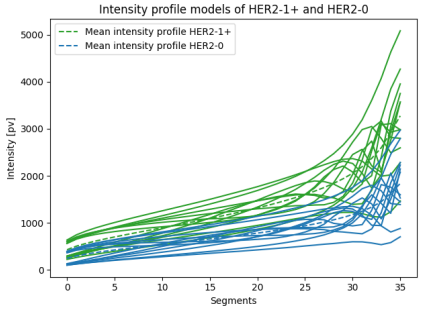
The different intensity profile models, separated by HER2-class, of the cells in the CPA:s in the training set is presented below. On the far left we see HER2-3+ and HER2-2+, in the middle HER2-2+ and HER2-1+ and to the right HER2-1+ and HER2-0. The different modes of variation that explains 99 % of the variation are presented for each profile and decides the width of the plot. Keep in mind that the segments have been sorted to remove spatial dependency. In all figures there is separation between the different intensity profile models and they are in general different regarding linearity and magnitude. There is a small overlap between the HER2-1+ and HER2-0 profile models.



(a) HER2-3+ and HER2-2+



(b) HER2-2+ and HER2-1+



(c) HER2-1+ and HER2-0

Figure 37: Intensity profile models of the different classes

4.3.1 Prediction on CPA:s

Using the ratios over the entire sample or over the ROI:s will be the same for the CPA:s as the CPA:s are considered as one ROI. In Table 6 the thresholds that decided the sample classification are presented.

Table 6: Thresholds that optimised the respective confusion matrix for the training set

Threshold	Cell ratio	Absolute quantity
3+	0.08	1600
2+	0.02	300
1+	0.38	7800

The largest threshold is between HER2-0 and HER2-1+ as this is the most difficult distinction.

Below the confusion matrices for both methods, using an absolute number of cells and using ratios to deduct the sample classification, on the training set are presented:

	3+	2+	1+	0
3+	32	0	0	0
2+	0	32	0	0
1+	0	0	30	2
0	0	0	1	31

(a) Cell ratio (97.7%)

	3+	2+	1+	0
3+	32	0	0	0
2+	0	32	0	0
1+	0	0	26	6
0	0	0	2	30

(b) Absolute quantity (93.8%)

Figure 38: Confusion matrices on the training set

The confusion matrices above describes that the model using the cell-ratio as the threshold type is better at HER2-0 and HER2-1+ distinction than the model using absolute quantity as threshold type.

The following is the classification results for both methods on the CPA:s of the test set:

	3+	2+	1+	0
3+	11	0	0	0
2+	0	11	0	0
1+	0	0	11	0
0	0	0	0	11

(a) Cell ratio (100%)

	3+	2+	1+	0
3+	11	0	0	0
2+	0	11	0	0
1+	0	0	9	2
0	0	0	0	11

(b) Absolute quantity (95.5%)

Figure 39: Confusion matrices on the test set

The same pattern as the training set applies on the test set as well.

4.3.2 Prediction on Tissue using ROI-annotation

Below the results for the three methods are presented. Th three methods again being: Using the sample cell ratio, the mean ROI-cell-ratio and the absolute quantity of cells as the basis for sample classification. The following thresholds, in Figure 54, decided the sample classification. The *training set* column describes the thresholds deducted from training and the *All images* describes the thresholds deducted from using all images as a training set.

Threshold	Training set	All images
3+	0.02	0.02
2+	0.04	0.04
1+	0.20	0.21

(a) Sample ratio

Threshold	Training set	All images
3+	0.03	0.03
2+	0.04	0.04
1+	0.18	0.18

(b) Mean ROI-ratio

Threshold	Training set	All images
3+	4100	4000
2+	3600	3600
1+	2000	13000

(c) Absolute quantity

Figure 40: Thresholds that optimised confusion matrices for training set and all images respectively

Similar to the case of the CPA:s the threshold is largest for the HER2-0 and HER2-1+ distinction. There is also a significant increase, when using all images as a training set, of the 1+-threshold when looking at figure 40c using absolute quantity as the threshold type.

Classification result on the training set is presented below:

	3+	2+	1+	0
3+	8	0	1	0
2+	5	3	0	0
1+	0	0	8	0
0	0	0	4	4

(a) Sample ratio (69.7%)

	3+	2+	1+	0
3+	8	0	1	0
2+	3	5	0	0
1+	0	0	8	0
0	0	0	4	4

(b) Mean ROI-ratio (75.7%)

	3+	2+	1+	0
3+	7	2	0	0
2+	3	5	0	0
1+	0	0	8	0
0	0	0	7	1

(c) Absolute quantity (63.6%)

Figure 41: Confusion matrices on the training set

By looking at the figure above using ratios are again better at distinguishing between HER2-0 and HER2-1+. Using the mean ROI-cell ratio is slightly better at HER2-3+ prediction.

Classification result on the test set is presented below:

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	2	1

(a) Sample ratio (63.6%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	2	1

(b) Mean ROI-ratio (63.6%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	3	0
0	0	0	3	0

(c) Absolute quantity (63.6%)

Figure 42: Confusion matrices on the test set

The same pattern as for the training set repeats itself, however the ratio models misclassifies a HER2-1+ sample as a HER2-0 which the model using absolute quantity as a threshold type classifies correctly. This causes identical accuracy as the latter misclassifies all HER2-0 as HER2-1+ and the two former at least classifies one of these correctly.

If we further consider all the images as available training material and train over all these the following confusion matrices where deducted:

	3+	2+	1+	0
3+	9	0	1	0
2+	6	6	0	0
1+	0	0	10	1
0	0	0	6	5

(a) Sample ratio (68.2%)

	3+	2+	1+	0
3+	9	0	1	0
2+	4	8	0	0
1+	0	0	10	1
0	0	0	6	5

(b) Mean ROI-ratio (72.7%)

	3+	2+	1+	0
3+	8	2	0	0
2+	4	8	0	0
1+	0	0	11	0
0	0	0	9	2

(c) Absolute quantity (65.9%)

Figure 43: Confusion matrices on all images as training set

Using all the images as a training set causes the same pattern as previously mentioned. Using the mean ROI-cell ratio as a threshold type performs better than the models using the other threshold types.

4.3.3 Predictions on tissue using complete tissue annotation

The results in this section are deducted by using the complete tissue annotations, including every part of the tissue as ROI. This makes the average ROI-cell ratio and whole sample cell-ratio identical as there is only one ROI. See Appendix B for result using all images as training set with this annotation. The following thresholds decided the sample classification:

Table 7: Thresholds that optimised confusion matrices for training set

Threshold	Cell ratio	Absolute quantity
3+	0.06	14500
2+	0.04	8700
1+	0.17	59300

The thresholds behave similarly using complete tissue annotation as they have behaved with previous ROI-annotation.

The following are the confusion matrices on the training set

	3+	2+	1+	0
3+	7	1	1	0
2+	1	7	0	0
1+	0	0	8	0
0	0	0	4	4

(a) Cell ratio (78.7%)

	3+	2+	1+	0
3+	8	1	0	0
2+	2	6	0	0
1+	0	0	7	1
0	0	0	5	3

(b) Absolute quantity (72.7%)

Figure 44: Confusion matrices on the training set using complete tissue annotation

The model using cell-ratio is the model that, accuracy wise, performs best on the training set and is better at the HER2-0 and HER2-1+ distinction. Using absolute quantity as the threshold type seems to be slightly better at HER2-3+ prediction.

The following are the confusion matrices on the test set.

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	2	1
0	0	0	1	2

(a) Cell ratio (72.7%)

	3+	2+	1+	0
3+	1	0	0	0
2+	1	3	0	0
1+	0	0	3	0
0	0	0	2	1

(b) Absolute quantity (72.7%)

Figure 45: Confusion matrices on the training set using complete tissue annotation

The two models perform equally good in the terms of accuracy on the test set and behave slightly different in the HER2-1+ and HER2-0 distinction.

4.4 Concatenated results by model

Below the training set accuracy is presented for each model. *Accuracy exp* describes the accuracy against the labelling by the pathologist experienced with UCNP-images and *Accuracy 1 correct* describes the accuracy if allowing the model to align with at least one of the pathologists. Refer to Table 3 for abbreviation explanation.

All the models have similar accuracies on the training set and if aligned to one of the pathologist they all perform with high accuracy as can be seen in Table 9. Studying Table 9 the multi-dimensional Gaussian model performs best on the test set and if aligned to one of the pathologists it performs perfectly. The far worst model is the Naive Gaussian model using ratio as threshold type.

Table 8: Accuracy table for the models on training set

Model	Accuracy Exp [%]	Accuracy 1 correct [%]
NGM 1	69.7	87.9
NGM 2	72.7	90.9
MDM 1	75.7	100.0
MDM 2	69.7	90.9
IPC 1	69.7	96.9
IPC 2	75.7	96.9
IPC 3	63.6	93.9

Table 9: Accuracy table for the models on the test set

Model	Accuracy Exp [%]	Accuracy 1 correct [%]
NGM 1	72.7	90.9
NGM 2	45.5	81.8
MDM 1	81.8	100
MDM 2	63.6	90.9
IPC 1	63.6	90.9
IPC 2	63.6	90.9
IPC 3	63.6	100.0

4.4.1 HER2-1+ and HER2-0 grouped together

Now consider a result where HER2-0 and HER2-1+ is regarded as the same category (negative). Grouping HER2-0 and HER2-1+ as one category increases the performance significantly for all the models on the training set, Table 10. Most significant increase is seen for the IPC-models. Accuracy is increased for the test set as well, Table 11, and now multiple models performs on par with each other and performs perfectly when aligned with one of the pathologists.

Table 10: Accuracy table for the models on the training set (HER2-1+/0 grouped)

Model	Accuracy Exp [%]	Accuracy 1 correct [%]
NGM 1	84.8	93.9
NGM 2	87.8	93.9
MDM 1	90.9	100.0
MDM 2	81.8	93.9
IPC 1	84.8	96.9
IPC 2	87.8	96.9
IPC 3	84.8	96.9

Table 11: Accuracy table for the models on the test set (HER2-1+/0 grouped

Model	Accuracy Exp [%]	Accuracy 1 correct [%]
NGM 1	81.8	90.9
NGM 2	81.8	90.9
MDM 1	90.9	100.0
MDM 2	90.9	100.0
IPC 1	90.9	100.0
IPC 2	90.9	100.0
IPC 3	90.9	100.0

4.4.2 Using all images as training material

Now consider all images as training material and the accuracy Table 12. Using all images as training material increases the accuracy significantly for the Gaussian models in contrast to the IPC-models.

Table 12: Accuracy table for the models using all of the images as training material

Model	Accuracy Exp [%]	Accuracy 1 correct [%]
NGM 1	86.4	95.5
NGM 2	86.4	93.2
MDM 1	90.9	100.0
MDM 2	84.1	93.2
IPC 1	68.2	95.4
IPC 2	72.7	95.4
IPC 3	65.9	95.4

4.4.3 Combined models

Looking at the two most prominent models and combining their score for both the training set and the test set the accuracy towards the experienced pathologist is shown in Table 13. The accuracy is increased on the training set but not on the test set when combining the models.

Table 13: Accuracy table for the individual models as well as the combined models

Set	Combined models [%]	IPC-2 [%]	MDM-1 [%]
Training set	78.7	75.7	75.7
Test set	81.8	63.6	81.8

4.4.4 Results on cross-validation

When computing a k-folds cross-validation, with k = 4, for the complete sample set, accuracies according to Table 14 were found. The IPC-model using mean ROI-cell ratio is the model that performs best over different folds and seems to be the most stable model. The accuracy of the Gaussian models fluctuates heavily between different folds. The same pattern as described for the accuracies above applies on the thresholds, Table 15. The IPC-model using mean ROI-cell ratio is the model that have the most stable thresholds whereas the other models have quite substantial fluctuations.

Table 14: Accuracies for the models using 4 different training and test sets

Model	Acc 1 [%]	Acc 2 [%]	Acc 3 [%]	Acc 4 [%]	Mean of Acc [%]
NGM 1	54.5	72.7	81.2	54.5	65.9
NGM 2	54.5	54.5	63.6	54.5	56.8
MDM 1	54.5	72.7	90.9	63.6	70.5
MDM 2	54.5	54.5	72.7	54.5	59.1
IPC 1	36.5	54.5	54.5	63.6	52.7
IPC 2	54.5	63.6	63.6	72.7	63.6
IPC 3	63.6	54.5	63.6	45.4	56.2

Table 15: Threshold ranges for the different models using the folds

Model	Th3+ range	Th2+ range	Th1+ range
NGM 1	8 - 12	40 - 80	75 - 120
NGM 2	0.01 - 0.02	0.06 - 0.31	0.0 - 0.38
MDM 1	4 - 11	80 - 180	75 - 90
MDM 2	0.01 - 0.02	0.10 - 0.31	0.05 - 0.52
IPC 1	0.02 - 0.05	0.02 - 0.04	0.19 - 0.48
IPC 2	0.03 - 0.05	0.04 - 0.04	0.18 - 0.43
IPC 3	4000 - 26000	2000 - 4000	13000 - 24000

4.5 Annotations on complete tissue

Below a table showing the results on the different sets using different annotation types for the most prominent models is shown.

Table 16: Results for the best models on complete tissue annotations

Set/annotation type	MDM 1 [%]	IPC 2 [%]
Training set/ROI	75.6	75.6
Training set/Complete	72.7	78.7
Test set/ROI	81.8	63.6
Test set/Complete	81.8	72.7

For the models above, using complete tissue annotation either increases or delivers the same accuracy as using ROI-annotation.

4.6 Comparison with previous work

Below the Table 17 is presented where the accuracy of existing algorithms in the field is compared with the best developed models. The accuracies (agreement with manual classification) for the existing algorithms and models are not evaluated on the same set and are evaluated on traditional IHC-images. A comment column describes if anything should be noted for the model. The developed models perform similarly in comparison with other existing models in the field except for the neural network model Her2net that exhibits far superior accuracy.

Table 17: Overall comparison with previous methods

Model	Accuracy [%]	Comment
Tissue IA	90.9	0 and 1+ grouped together
Her2net	99.05	Sample prediction unclear & using 40X magnification
VisioPharm Her2CONNECT	91.8	0 and 1+ grouped together
IPC-best	90.9	0 and 1+ grouped together
MDM-best	90.9	0 and 1+ grouped together

4.7 Alignment

Below the Table 18 is presented where the alignment between the three pathologists is shown as well as alignment with the best of the developed algorithms. The table is based on the test set for the algorithms and the complete set for the pathologists. The models align quite well with all pathologists and more often than not aligns better than the pathologists align with each other.

Table 18: Alignment table

	Pathologist 1 (exp)	Pathologist 2	Pathologist 3	IPC-2	MDM-1
Pathologist 1 (exp)	100%	75.0%	54.5 %	63.6 %	81.8%
Pathologist 2	75.0%	100%	59.1 %	81.8 %	81.8%
Pathologist 3	54.5%	59.1%	100 %	72.7 %	72.7%
IPC-2	63.6%	81.8%	72.7 %	100 %	81.8%
MDM-1	81.8%	81.8%	72.7 %	81.8 %	100%

5 Discussion

5.1 Annotations

The annotations of the samples were not made by pathologists which naturally causes some concern. The annotations were based on the criteria in section 3.3.1, and those judgements were not made by specialist in the field. This thesis shows that it is possible to classify UCNP images by using interpretable image analysis but the thesis would be strengthened further by including pathologists in the process as the annotations then would follow field standard. This was not made due to lack of resources and time. There was also a reasoning that the classifying methods should be robust to annotations. And as the results shown in Table 16 the classifying methods are robust to annotations since including all the tissue in the annotation results in similar or better classification accuracy. It is likely that many misclassifications and erroneous results when using ROI-annotation is due to faulty annotations. Using annotations from experts in the field would likely improve the results.

5.2 Cell detection in QuPath

The cell-detection by QuPath was occasionally slightly erroneous which resulted in a membrane segmentations that were rather a good approximation than a truth. However, the classifying methods should be robust to wrongful segmentations and the number of cells with a decent cell segmentation often outweighed the number of cells with a poor cell segmentation. We would, however, expect that performance would improve with a better cell detection algorithm since this would yield more accurate features.

5.3 Test set distribution

The test set is overly weighted towards HER2-2+, HER2-1+, and HER2-0. There are 3 examples of the two latter and four of the former and only one example of HER2-3+. This is not ideal as the models are better at predicting HER2-3+ and this should be the most important to predict as well. The test set distribution was not investigated on forehand based on the labels of the pathologists but instead of those from the sample manufacturer. Knowledge was later acquired that these labels were not particularly accurate and instead the labels of the most experienced pathologist were used. This caused a skewed test set distribution. For future work the test set distribution should be properly investigated. It would also be advantageous to work with a larger test set to further evaluate the results. A more suiting dataset size would consist of 100 images where 20 images would be the test set. WSI:s consist of a large amount of data so necessarily one would not need too many images and as access to images derived from healthcare might be restricted one might have to be satisfied with smaller datasets.

5.4 Edge effect

In some samples an edge effect was visually detected, and this is not unique for using UCNPs but is a normal effect in IHC. This means that there is more conjugate in relation to cells along the edges of a sample than in the centre. This means that the edges will appear brighter and appear as they have more expression when in reality that might not be the case. This has an impact on the results and could cause overclassification if the edges have this property and if the edges are included in the annotation. This can be seen in Appendix A.3. Naturally the edges could be excluded if the annotation is done manually. If the annotation is done automatically one would have to, after automatic segmentation of the tissue in the image, integrate a cropping algorithm that excludes a suiting amount of the edge.

5.5 Training on HER2-tissue samples

Ideally and in classical image classification theory one usually works from a set of labelled images where one could extract features corresponding to those labels. This is not a particularly good approach when working with HER2-tissue samples. Within a sample there

can be big subregions that does not exhibit any HER2-expression which, if the sample is big, would overall create a bias towards lower classes even though the sample pathologically should be classified as a higher class. Ideally, if one has access to a pathologist resource (ideally several) one could use annotations made from the pathologist where a score is given for each annotation as training data. This does, however, require a general consensus of the annotation score between the assessors and a substantial amount of manual labour. When that data is not available a good approach would be to match subregions to the CPA:s and determine the sample class from there, either by choosing the maximal class of the subregions or by following some sort of methodology that can be correlated to ASCO-guidelines (for instance 10 % of the total ROI-area should be 3+ or similar). Naturally a CPA will differ from real tissue as these cells are cultivate and not as diverse as real tissue cells which will cause some classification difficulties.

5.6 Bayesian models

It is clear from Figure 17 - 19, that there is a large difference between HER2-3+ and HER2-2+, HER2-2+ and HER2-1+ for all features included in the model. As the graphs displays their probability density functions are clearly separated, which indicates high class difference. That the classes are separable based on the features used is of course important for the Gaussian model, since overlapping probability density functions will lead to a high risk of misclassification. In other words, an unknown sample will have a high probability of belonging to the another class if probability density functions are overlapping. While examining Figure 17-19, it is observable that the probability density functions of the features for HER2-1+ and HER2-0 are overlapping. Given the training set, the features that were extracted were not statistically separable and there is therefore a high risk of misclassification between HER2-1+ and HER2-0.

From Figure 20 it is clear that all features combined from HER2-3+ and HER2-2+ are separable to HER2-1+ and HER2-0. However, these features did not in a combined way manage to separate HER2-0 and HER2-1+. These separation issues are probably the cause of many misclassifications in the test set for CPA:s and tissue samples.

5.6.1 Prediction on CPA

In section 4.2.1 the trained classifier predicts CPA:s from the training set. Figure 38 a), shows the confusion matrix after prediction with a Bayesian classifier using only normalised intensity as feature. With an accuracy of 93.75% it can be concluded that this feature performs very well on its own, while predicting on the training data. This confirms our hypothesis that the intensity value in an UCNP image will be highly correlated to the class. The results were slightly improved, from 93.75% accuracy to 95.3% accuracy by adding two more features to the model, see Figure 38 b). A multidimensional model was hypothesised to mitigate the diversity between the cultivated cells and the tissue cells and therefore have a higher chance of finding the correct class.

Both the Bayesian classifiers using only normalised intensity and the multidimensional classifier performed with the same accuracy, 97.7% on the CPA:s in the test set. With this accuracy on both models, it can be concluded that the models are robust and trustworthy if the images trained upon and tested upon are similar to one another. This is a necessity for the classification models to classify tissue images correctly as performed in a later stage. If the models would perform poorly on CPA:s while trained on CPA:s, they would most likely perform even worse while predicting tissue, since the cultivated cells and real tissue cells have differences in their appearances. Both classifiers performed too good to be discarded when moving on to predict tissue.

5.6.2 Prediction on tissue training set

The results on tissue was not as promising as the results on CPA:s. The accuracy was between 70-76% for all models and thresholds investigated. The main reason for why these results were not as strong as the results on the CPA is the diversity of a cell in tissue,

whereas the CPA:s are cultivated cells that are very similar to one another. The model was also trained on CPA:s and evaluated on tissue, a methodology that not is ideal as discussed in previous sections. A high difference between the training data and the validation or test data will of course amplify risks of misclassification, since a feature value found in validation or test sample might not be close to any of the feature values in the training data. The model simply finds the best match of the class, not necessarily a good match.

5.6.3 Prediction on tissue in test set

While using the threshold values found during training to classify tissues from the test set the accuracy for some models stayed approximately at the same level as during training, while others dropped significantly. Using ratio for Bayesian classifier with normalised intensity as the only features performed with the worst accuracy, 45.5 %, and the multidimensional using an absolute quantity of patches performed with best accuracy, 81.2%. The confusion matrices in Figure 27 and Figure 28 displays how all models are performing poorly in differentiating between HER2-1+ and HER2-0. These results were expected since the images in classes HER2-1+ and HER2-0 are very comparable. There are very few features that differentiate one from the other. Misclassification of the patches could be the problem, which makes the quantity or ratio of patches needed for each class very uncertain. Another reason could also be that the threshold values are not optimised since they could only be deducted from 33 images. A larger training set would probably contribute to better predictions in the test set.

5.6.4 Prediction on test set using complete annotations

Using complete annotations the accuracy improved on the test set for all models except the multidimensional Bayesian classifier using absolute quantity as threshold type. No unwanted distribution in the confusion matrices were observed when switching from ROI-annotations to complete annotations, it only failed between HER2-3+ and HER2-2+, and between HER2-1+ and HER2-0. It can therefore be concluded that the models perform better if more information and data is given to them. Despite adding irregularities or parts of data that not is manageable for the model to work with, the perks of more data seem to be larger than being restrictive with regions of interest. It could also be concluded that the threshold models are robust to changes of the annotated regions.

5.6.5 Cross fold validation on Bayesian models

To evaluate the models robustness to different training and test sets a cross fold validation was completed on the ROI-annotated images. As displayed in Table 14 the models that occasionally perform well have a a high variation of accuracy for each fold. It can therefore be concluded that the robustness of the Bayesian models not are very high, and that they are highly dependent on the training and test set division. The models that performed with the highest accuracy on the original test set for ratio annotation, see table 27 & 28, continued to perform well in some of the folds, whereas their accuracy drop significantly in other folds. The reason for this drop can be many. One of the reason for these results could be that the training and test set not are very similar in these folds. Another reason could be a skewed distribution of classes in the training or test set, teaching the model to learn well on some classes but it is being tested on other classes than trained on.

5.6.6 Hyperparameters

The patch size was chosen as the default tile size of the WSI-format of Lumito to optimise file-compatibility and to include a sufficient amount of cells. The sufficient amount of cells was set to 50 and what is sufficient can be discussed and there is room for fine tuning this parameter to improve performance even though we expect it to be of negligible importance. One could optimise the patch size towards performance on the training set and for future work this should be investigated between the ranges of 100 and 1000 pixels dependent on the annotations. This could very well impact performance significantly.

5.7 Intensity profile classifier

The top 5 % pixel values of a membrane-segment where used to create an intensity profile classifier as this produced the smallest amount of misclassifications which can be seen by looking at table 5. This evaluation was made using minimising misclassification of cells as the optimisation criteria and not maximising classification performance over the tissue samples in the training set. It is possible that one could use this criteria instead to achieve better performance, it is somewhat trickier and not as time effective as the evaluation used. It is furthermore not very probable to significantly increase performance.

As one can see by looking at figure 36 the intensity profile model of all the cells vary mostly in two ways. Primarily it is the magnitude of the profile that varies and that describes how much intensity that has been found within the membrane segmentation, more specifically how much expression that exists on the cell membrane. Secondly it is the slope of the profile that varies and that would describe how well covered the membrane is by the HER2-receptor. Ideally a HER2-3+ cell would have a flat line with high intensity and where we would expect a more linear pattern in HER2-2+. But because the cell-segmentation is somewhat of an approximation the behaviour in a 3+ would be more linear than flat since there could be multiple segments with low or no expression. These two modes of variation correlates directly to the ASCO-guidelines and would be interpretable for a pathologist.

By further looking at the sub figures in figure 37 we can see that there is a clear difference in both magnitude and shape between the HER2-3+ profile model and the HER2-2+ profile model. Shape wise, there are only a few segments with high intensity values in the HER2-2+ profile whereas the distribution is more even in the HER2-3+ profile model. Magnitude wise, the intensity is higher overall in the HER2-3+ profile model. The difference is similar in the HER2-2+ and HER2-1+ plot, the shape is in both profile models similar but the 2+ has a steeper ascent which could be explained by the fact that the 2+ has a higher intensity difference between the segments overall. Magnitude wise it is a clear separation in intensity between the two profile models. Moving on to study the HER2-1+ profile model in comparison to HER2-0 they are very similar in shape but differs slightly in magnitude. There is a slight overlap between the two profile models but yet a clear separation.

5.7.1 Thresholds

The largest threshold, for CPA prediction, is the threshold for the HER2-0 and HER2-1+ distinction which can be seen by looking at table 6. This is because this is the most difficult distinction, it is very probable that there are a significant amount of HER2-1+-cells in a HER2-0-CPA-pellet. The other thresholds are lower as there usually are a fewer number of HER2-3+ cells in the HER2-2+-CPA-pellets and the same applies for HER-2+ cells in HER2-1+ CPA-pellets. For example, for the HER2-3+ and HER2-2+ distinction, once the threshold of 0.08 is found the model can differentiate between all HER2-2+ CPA-pellets and HER2-3+ CPA-pellets since no HER2-2+ CPA-pellet contains more than 8 % HER2-3+cells.

Looking at the thresholds in figure 54 one can see the same pattern in tissue noticed for the CPA:s. The highest threshold is again between HER2-0 and HER2-1+, for the same reason. It is high chance of HER2-1+cells occurring in a HER2-0 sample and vice versa. The low magnitude of the other thresholds however are not because of the same reason as for the case of the CPA:s. Those thresholds are rather low because that in a HER2-3+-sample the overall occurrence of HER2-3+-cells can be, in relation to the lower categories, quite low as there are cells carrying no or little HER2-expression. And the same applies for the HER2-2+case as well. Because of this the thresholds need to be quite low in order to find those cases where this happens.

5.7.2 Prediction on CPA:s

As can be seen by looking at the results in 4.3.1 the intensity profile classifier works extremely well while working with CPA:s. And when predicting CPA:s on the test set there is a 100 % accuracy for the model using cell-ratios in the different categories as threshold type and a 95.5 % accuracy when using the absolute quantity of cells in the different categories as the

threshold type. Both models performed acceptable on the CPA:s in the training set which validated to move forward with both models to predicting on tissue.

5.7.3 Prediction on tissue

Moving on to tissue, and observing the results in section 4.3.2, the three models behave similarly on both test and training set, the main difficulty is distinguishing between HER2-0 and HER2-1+. The model using absolute quantity as threshold type has the most difficulties in this distinction. The model using average ROI-cell-ratio as the threshold type is the model that is the best when classifying HER2-3+-samples which makes it slightly more prominent than the others. In terms of accuracy on the test set, however, the difference is none with the exact same accuracy (63.6 %). There is a larger difference in the training set where the model using absolute quantity as the threshold type performs far worse than the other two. The ratio models have accuracies ranging from 70 - 75 % and the absolute quantity models has an accuracy of 63.6 %.

5.7.4 Prediction using complete annotations

By looking at the results in section 4.3.3 one can see that; by using the complete tissue annotation the difficulties in distinguishing between HER2-0 and HER2-1+ persists even though performance in this regard is increased slightly. The logic behind this could be that areas with no or little expression are included more with this type of classification which would make it easier to distinguish between the two lower classes. The same applies for HER2-2+ and HER2-3+ distinction as this is also improves with using complete annotation. Both models perform similarly when using complete tissue annotations. Using absolute quantity as the threshold type seems to lead to a higher risk for misclassification between HER2-0 and HER2-1+ but also to a slightly higher leniency regarding HER2-3+ prediction. In general the accuracies increase using complete tissue annotation which concludes that the intensity profile classifier works best when including an entire tissue sample as the prediction information. This, regardless of what threshold type is being used.

5.7.5 General conclusions for the IPC

Studying the tables in section 4.4 one will observe that the best intensity profile classifier is the one using the average ROI-cell-ratio in the sample to decide the sample classification. At least if one considers accuracy over both training and test set. And by observing the k-fold cross validation results in table 14 and 15 one can observe more threshold stability as well as stable performance over different data-sets for this model in contrast to the other two models. This seems to be in line with the appearance of a HER2-tissue sample where small parts of a tissue sample can exhibit high HER2-expression and subsequently could decide the sample score. Using total sample cell-ratio one would expect that small and high-expression parts would be averaged out over the sample whereas the average ROI-cell-ratio would better reflect these. The absolute quantity intensity profile classifier performs similarly to the other models based on ratios but will have difficulties to set an absolute threshold due to varying amount of cells in the samples. This is reflected in the figure 54 where including all of the test images in the training set significantly changes the threshold used to distinguish between HER2-0 and HER2-1+. Using ratios, in general, also produces a better result on the CPA:s and seems intuitively more stable as the size of the sample is neglected in this way. When using complete tissue annotation there are small but significant differences between using quantities or ratios as the threshold type in the favour of the latter.

The intensity profile classifier was trained and based on cells from the CPA:s. This will make it difficult to predict cells in tissue perfectly, as previously mentioned. The results from the intensity profile classifier reflects in a small way how far you can get by using the CPA:s as training data. As the cells differ slightly there can never be optimal performance on tissue (even though the performance is rather good). If one would have access to a data set with annotated cells in tissue in the different categories the performance on tissue is likely to improve. This would however require a significant amount of annotation-work to build up reliable intensity profile models with similar amount of data behind it as the ones

in this report. It is also probable that a better cell detection algorithm would improve the results as the characteristics of the different HER2-categories then would be captured more accurately.

5.8 Comparison between the models

The two models are very similar. Both models usually extract information on a cellular level where the IPC use it directly to create intensity profile models for the different HER2-categories and where the Bayesian model use the information to build Gaussian distributions for the different HER2-categories. The IPC only needs one feature to work and is rather simple to train. The Bayesian model can vary in simplicity using one or more features. The normalised intensity feature gives a lot of information and does not require information on a cellular level (only the amount of cells in a specific area). Adding more features to the Bayesian model requires cellular information. Both models require a set of thresholds to predict on a sample level. Either an absolute quantity of cells/patches in the different HER2-categories or some sort of ratio of the same. The main difference between the models are that the Bayesian model works patch wise and the IPC-model works cell wise. The IPC-model calculates a mathematical distance from a cell to each intensity profile model whereas the Bayesian model calculates a statistical distance for each patch to each Gaussian distribution. Visually the two models perform similarly which could be seen in the images in Figure 13-16, the Bayesian model tends to, however, classify areas a bit higher than the IPC-model. The visual similarity is expected as the models both use the feature that describes the membrane coverage where the Bayesian models (the multidimensional one) use it on a higher level.

By observing Table 8 the models seem to perform rather similarly. However, moving on to study Table 9 the Bayesian models seem to outperform the IPC-models. Both models struggle to to distinguish between HER2-0 and HER-1+ which is shown in Table 10 where accuracies on the training set increase considering HER2-0 and HER2-1+ as the same class. The Bayesian models seem to be slightly better at this distinction which is further strengthened by Table 11, where the accuracies are on par with each other, which evidently causes the difference in performance on the test set. This could be because that the IPC, as previously discussed, could be overly adjusted to the CPA:s whereas the Bayesian models extract similar information but on a higher level that is more applicable on tissue. Observing the k-fold-cross validation Table 14 however, the intensity profile classifier (IPC-2) performs more stable. This could be used to argue that the IPC is more robust over different datasets and that the Bayesian models are more dependent on balanced training sets. However, the stability could also be a result of that the IPC-model simply can not reach the same high accuracy values as the Bayesian models. It would be further interesting to compare this result to when the different folds are forced to be balanced.

It is unclear, at this stage, if one method is better than the other. Further validation over a larger dataset would probably determine that. Interestingly the Bayesian model that performs best is the multidimensional model using an absolute quantity type threshold as a basis for classification where we do not see that behaviour in the IPC-model. At this stage it is unclear if this holds over a larger data set and also unclear why the two models differ in this instance. Both models are, as previously discussed, robust towards annotations and could also increase in performance if the entire tissue is included in the annotation. The fastest of the models is the Bayesian classifier only using normalised intensity as this method does not use any features derived from the cell-segmentation. This model is the one that primarily should be considered for simple applications. All the other models are comparable regarding speed.

5.9 Impact of pathology labelling

As can be shown in the Table 18 the pathologists on these samples seldom agree. And if the accuracy were derived with the basis that they should agree with at least one of the pathologist one can observe by looking at the tables in section 4.4 that the models perform very well. Allowing the models to agree with only one pathologist gives a result that is pow-

erful which leads to deduction that the models performs comparably to a pathologist. By again observing table 18 the algorithms more often than not aligns better than the pathologists align with each other which further strengthens the result that the models perform comparably to a pathologist. One should keep in mind, however, that two of the pathologists were unfamiliar with assessing UCNP-images which would cause a bigger spread in the HER2-assessment than normal. This is especially the case for the pathologist 3 as can be seen in the table 18. This downplays the result that the models perform comparably to a pathologist and it is unclear how well this holds for when comparing with, for example, five pathologists familiar with the UCNP-images.

The pathologists in general performed reasonably and in line with general intensity and coverage values. In some cases the expertise seemed to extend beyond those features and in the Appendix A.4 there are some cases of where the algorithms behaved correctly but where another, to the authors unknown, factor decided the score.

5.10 Deduction of threshold values

The threshold values were deducted using the labelling of the pathologist with most experience with UCNP-images. This is not necessarily correct and naturally creates classifiers with bias towards to this particular pathologist. Instead one could deduct thresholds based on the labelling for each pathologist and concatenate these by averaging or similarly to remove pathologist bias. This would probably produce better results on unseen data. In this case however, due to the state of the art technology the experience with UCNP-images was greatly valued in the sense that the familiarity with the image format would lead to a more confident diagnosis. And this outweighed the drawback of a bias.

5.11 Comparison with previous work

By looking at table 17 we can clearly see that the developed algorithms are on,regarding accuracy, par with existing methods that use traditional image analysis on IHC-images like SlidePath's Tissue IA and Visiopharm's Her2CONNECT but fail to deliver the same accuracy as a neural network. The existing methods have not been evaluated on UCNP-images and it is unclear how well these would perform using UCNP-images instead of traditional IHC-images.

5.11.1 Comparing with traditional image analysis methods

What advocates for the developed algorithms in contrast to the existing traditional image analysis models are the lack of dependencies towards colour/staining consistency. UCNP staining seems more stable and produce one channel HER2-expression separated images which leaves less room for variance in comparison with traditional IHC-images. The developed algorithms rely solely on the thresholds deducted from training such as the ones presented in table 54 and they seem to be, for most of the models, rather stable even when including the test set in the training set except for the 1+-threshold but as the 1 and 0 classification is grouped together for this comparison (refer again to table 17), this is not applicable.

Advocating for the existing methods, the existing methods do not seem to require as extensive pre-computation of the data whereas the developed (the more refined ones) methods rely on cell-segmentation algorithms and use extensive computations based on these segmentations. The developed algorithms perform well with complete annotations and would only need an outline of where the tissue is situated to classify a sample. This means that these algorithms could be integrated with automatic ROI-algorithms. It is for the existing methods unclear what level of ROI-annotation they need to work properly but some annotation is required (at the very least overall-tissue-annotation) and Tissue IA seem to require more extensive annotation than HER2-CONNECT.

5.11.2 Comparing with neural networks

Comparing the developed algorithms to Her2net the latter is far superior when predicting on HER2-tissue. However there are a few bullet-points worthy of discussion. The first is the sample prediction of her2net, in the article by Monjoy Saha and Chandan Chakraborty there is no specification on how her2net classifies a WSI based on the patch predictions or if the accuracy simply is determined patch wise. Furthermore the patches are created from an already cropped out tile (2048x2028) from the WSI. Thus, it is even more unclear if they at all predicted the tile or only the patch. If the accuracy is determined patch wise they must have worked under the assumption that all patches of a breast cancer tissue sample classified with a HER2-score corresponds to that HER2-score, which is a big and incorrect assumption, or have had access to a pathologist resource that determined the ground truth patch wise. Most likely it is the latter case but there is no description of how that was done. If they do determine the accuracy patch wise it is unclear how well that accuracy would transfer to tile-prediction and further WSI-prediction accuracy. This makes the accuracy from the developed models quite difficult to compare to the accuracy of her2net.

The second discussion point is that the WSI:s used for her2net are obtained with a magnification of 40X which means that there is four times as much information per tile in comparison with the WSI:s used in this work. This further makes the results incomparable and it would be interesting to see how well her2net performed on set with 20X WSI:s.

The third and final point of discussion is the complexity and interpretability of her2net. It is not very easy to explain the network structure and the basis for classification to a pathologist and there is no clear correlation to the ASCO-guidelines which can make this, as with many machine learning based diagnostic tools, less usable as there are no clear motivations behind each classification. Additionally if a pathologist still would have to assess the more important cases then there might not be much to win with extreme accuracy as a trade of with interpretability. To make the her2net useful for the clinic the issues regarding interpretability and tissue sample prediction would have to be solved. However, the performance of her2net is nonetheless impressive and it further strengthens what could be achieved with a neural network. It is of special interest that the separation between HER2-0 and HER2-1+ is accurate which can not be seen in the other models (see Table 17). Furthermore, execution time wise her2net (once trained) would also probably predict a WSI faster than the developed algorithm as one patch prediction took approximately one second whereas the developed algorithms require substantial pre-computations. Speed is always of importance from the clinical perspective.

5.12 Application in clinical practice

The results provided in this report show that a simple classifier works rather well on the UCNP-images and even though well refined classifiers were investigated the most simple ones provided acceptable and powerful results. And as the focus of this report was on simplicity and interpretability the simple classifiers are the ones that are, from our side, deemed to be the most successful. In the clinical perspective these simple classifiers could provide some screening assistance for the pathologists and decide which samples that are unnecessary to examine and which require more careful investigation.

The importance of how the models misclassifies then becomes a key factor, since a misclassification could lead to a wrongly stated diagnosis and treatment plan. The four classes could be grouped to two groups, HER2-3+ and HER2-2+ as group 1, and HER2-1+ and HER2-0, as group 2. As mentioned in the background group 2 will correspond to a negative HER2 diagnosis, since both HER2-1+ and HER2-0 will feed into the same diagnosis, therefore classification between HER2-1+ and HER2-0 would not be the highest priority in clinical practice. However, if the screening would aim to remove group 2 automatically and only use pathologists to examine samples from group 1, the differentiation between group 1 and 2 would be very important. Consequently, the confusion matrices created should have very few values in the four lower left parts and in the four upper right parts. Examining our confusion matrices from all models , one will find that these slots are usually empty. In

fact, most of the models presented have a 100 % separation between group 1 and group 2 over both training and test set whereas all the models have 100 % separation on the test set. That means that our models would suit well as a base for screening within pathology. As stated in the background only 12-15 % of breast cancers are HER2 positive. By removing all certainly negative HER2 samples pathologists would only need to distinguish between if the sample is positive or if it is equivocal, which would require less resources and free up time for a pathologist.

Because of large computing times when using intensity profile classifier and multidimensional Bayesian classifiers, the proposed model for a screening at this stage would be the Bayesian classifier that only uses normalised intensity as a feature. This model is only in need of the summarised intensity over the annotated tissue and the number of cells in that tissue. This requires less computational power than finding granular information around each cell within the tissue, that the other models are in need of. As Figure 34 expresses, it is capable of predicting the difference between HER2-3+/2+(group 1) and HER2-1+/0(group 2). In that way, this model can be well suited to use in a screening application.

In order for the screening to work one would have to implement a support for automatically finding ROI:s since the pathologist can not spend time to mark these for the algorithms. The main selling point would be that the pathologist is presented with the result of the algorithm and is being given a support system to find "hot-spots" of where to begin examination. If the pathologist would have to mark the ROI:s then no time is won, then they can assess the tissue already at that stage. For future work this would have to be integrated with an automatic ROI-detector and furthermore integrated with a viewer system such as QuPath or similarly. As can be seen by looking at the table 16 the algorithms perform well using complete tissue annotations. This would make them easy to integrate with an automatic ROI-detector that only detects tissue as these ROI-detectors are widely available. That result (in table 16) also shows that the algorithms are, as previously discussed, robust towards ROI-annotations which would mean that they could be integrated quite easily with any ROI-detection algorithm as briefly mentioned in 5.11.1.

The main takeaways are that there are some work needed for this to be implemented in an application but the possibility that it could be done persist where the classifiers presented could work as computer aided diagnostic systems using rather simple and interpretable mathematical methods.

5.12.1 Correlation to ASCO-guidelines and interpretability

Both of developed models can easily be correlated to the ASCO-guidelines, which was one of the main goals with the thesis. The intensity profile classifier is very correlated as the intensity profile, as mentioned, varies with both how much intensity there is around the cell and how the intensity varies around the cell. These both modes of variations describe the intensity of the staining and the completion of the staining around the membrane which is mentioned in Figure 1. The basis for classification of a cell can be described as how, mathematically, similar a cell is the general intensity profile for the different HER2-categories which a pathologist most likely would understand easily. The Bayesian classifier is also correlated to the ASCO-guidelines as it bases the classifications on features correlated to the amount of expression there is in relation to the amount of cells and features that describe average coverage or completion of staining in the area. The main difference between the classifiers is that the Bayesian classifier works on a higher level. The Bayesian classifier calculates a probability based on the features of an area for that area to belong to one of the HER2-categories. The biggest probability decides the class and entities such as probability and Gaussian distributions are entities that pathologists should be somewhat familiar with. However, none of the models make use of the thresholds correlated with the 10 % thresholds that can be seen in Figure 1. Initially these thresholds where considered as the ones to be used, but it became more complicated as these thresholds are not specified for a particular area of investigation. It could very well be very small ROI:s that are investigated as well as larger areas. This made these 10% thresholds quite difficult to work with.

5.12.2 Data extraction

For the models to work, each of WSI-images needed to be processed. The models could be easily trained on the CPA:s, where the cell-based features are derived to build the respective model. For each model, in order to predict a sample the same information needed to be found over the annotations in the tissue. For the Bayesian models data was extracted patch wise and for the intensity profile classifier models data was extracted cell wise. The extraction methods differed slightly but could be configured to extract overlapping data simultaneously in the future. This took some time for both models and for a data-set consisting of 44 images the extraction took approximately 15-30 hours dependent on the annotation-method applied. This is generally not acceptable and can not be used as an application. However, the data extraction method is not optimised and not sufficient time have been spent on investigating on how this could be tweaked and configured to work optimally to increase the speed. Furthermore, this was not the scope of the master thesis and there is work to be done to optimise these methods. An alternative is to run classifiers over night to save time where a pathologist could verify the results in the morning.

5.13 Future work

5.13.1 Prediction from cell or patch knowledge to tissue

The maximisation of the confusion matrix diagonal is one optimisation criteria to optimise the different thresholds against when moving from predictions of cells or patches to prediction of whole tissue samples. However, by only finding the maximum accuracy, the distribution between different individual classes sensitivity and specificity will not be equally distributed. Therefore, the thresholds found can have a chance of being weighted towards certain classes. An example of this is the multidimensional Bayesian classifier (33) and intensity profile classifier using absolute numbers (41) as the threshold type. In these predictions many true HER2-0 are classified as HER2-1+, and very few samples are classified as class HER2-0.

More robust methods could be used to find a more equal distribution of the misclassification in future work. An idea could be that the recall or precision for each class must be above a certain threshold, while simultaneously finding the best accuracy. In that way the algorithm will prioritise an equal distribution of correctly and incorrectly classified samples over all classes above finding the best possible accuracy.

5.13.2 Random forest

Using all the models derived above one could create a random forest classifier allowing the classifiers to either create a continuous score or vote for a score of a sample. A good thing with such a model is that the model would grow as more classifiers are developed. However, the samples that are misclassified by the models developed are usually the same but a random forest classifier could increase performance slightly as can be seen by the Table 13 where one model is allowed to be correct in alignment with the labels. This increase is quite small and only notable in the training set. It is unclear if this would improve performance but over a larger dataset but it could be worth investigating.

5.13.3 Classifiers independent from brightfield images

It could be of interest to develop a classifier that is independent from cell-segmentation algorithms and the brightfield image. This would create a classifier that could, without the knowledge of any cellular information, relying solely on the UCNP-image classify a tissue sample. By arbitrarily allowing an algorithm to divide a WSI into subregions and in those subregions find structures that one could correlate to ASCO-guidelines one could classify a tissue sample. For example, if a region exhibits many connected components with high intensity that could be a HER2-3+ region. This does however, still require annotations as the tissue and the CPA:s need to be marked. The latter because of consistent training material and the former because preventing unnecessarily big data. For future work we suggest that this should be investigated and we suggest that the information one should try to extract

would be the amount of connected components, their average size and intensity similar to the ones developed in the field for IHC-images. This should correlate well to the ASCO-guidelines as intensity quantifies amount of HER2-expression and connected components could be argued to describe mean coverage around the cells

5.13.4 More thorough cell assessment

The algorithm was developed under the assumption that all cells from a tumour sample are cancerous. Since it is only the cancerous cells that are assessed for the HER2-expression one would ideally sort out noncancerous cells from the tissue sample and from the cancerous cells only develop a classifier. This was nothing that was done and for future work one could try to sort out noncancerous cells by setting a minimum requirement of nucleus-size for example to catch the most simple noncancerous cells. This was not done in this report since the main scope was focused on the information deducted from the UCNP-images but naturally this could be combined with the cellular morphology information from the brightfield images.

6 Conclusions

From this master thesis we can conclude it is possible to classify UCNP-images by using interpretable image analysis methods in a way that is comparable with state of the art methods and current practice within the area. Grouping the two HER2-negative classes together the best methods perform with an accuracy of 90.9% and there is always a 100% separation between the HER2-negative and the other HER2-categories on the test set, even for the simplest of the algorithms. This indicates that the methods could be useful as a screening tool in the future. The algorithms would have to be validated and evaluated on a larger dataset to further validate the results. It is probable that the developed algorithms would increase in performance once combined with better cell-detection algorithms and introducing a pathologist resource to the development. Finally, there is a lot of promise of the technology with UCNP-images that Lumito provides. It is not only bypassing known difficulties within the IHC-field such as cell-morphology obstruction and staining consistency but also exhibits the possibility to develop interpretable classifying methods that performs similarly to existing classifying methods on IHC-images.

7 Ethical aspects

If introducing computer aided diagnostics into a clinic, it is important that it is implemented for the right reasons. It is of course lucrative for a clinic to use automatic screening, it would both reduce costs and increase efficiency considerably. However, this must only be implemented if the total benefits for all clients combined is increased or equal to previous diagnosis. It would be highly unethical to introduce such system only to benefit the clinics profits, without the performance of the whole system in mind.

It is also important how the algorithm it misclassifies. An algorithm that misclassifies many HER2-3+ as HER2-1+ might have a life threatening effect on the patient, whereas misclassification between HER2-1+ and HER2-0 would not impact the HER2 diagnosis significantly. Therefore it is important to not introduce a system that is not fully evaluated and confirmed that these misclassifications are better or equal to how a pathologist would perform.

Another ethical aspect worthy of discussion is if knowledge about images or specific types of structures that the algorithm not is able to classify correctly should hinder a computer aided system to be introduced or not. An algorithm is usually much more systematic in its false classifications compared to a human, who can be more random in their misclassifications. It would be important to discuss how special cases that the computer aided diagnostic system misclassifies should be handled and if a small proportion of samples should stop a system that besides these perform much better than a pathologist.

Lastly, even though a computer aided system performs with a worse accuracy than a pathologist it could be beneficial for all patients combined, since it would give pathologists more time to perform other tasks. The misclassified samples and therefore wrongly assessed treatment plans, due to the computer aided system might be less than the total new number of patients that a pathologist could help when more time is given to them. Of course this questions is up for debate, and neither answer is more correct than the other, but the ethical questions should be evaluated carefully before taking any decision in implementation of computer aided systems.

8 Contributions

Both authors have been involved in every process of the report but have had different responsibilities throughout the project. Alexander has been responsible for describing background related to breast cancer, the HER2-protein and the ASCO-guidelines whereas Adam has been responsible for writing about the Up-converting Nano-particle usage, Whole slide imaging and previous work. Adam has been mainly responsible for developing the intensity profile classifier models and Alexander has been in charge of the Gaussian models. The authors have with equal efforts contributed to the methodology and discussion. The conclusion is written by Adam whereas the ethitical discussion has been written by Alexander.

References

[1] Abcam. DAB staining. *Abcam*. https://www.abcam.com/kits/dab-staining. (Fetched 2022-02-01)

[2] Abcam. Immunohistochemistry(IHC) the complete guide.*Abcam*. https://www.abcam.com/content/immunohistochemistry-the-complete-guide. (Fetched 2022-02-01)

[3] American Cancer Society. Beast Cancer Facts Figures 2019-2020. *Breastcancer org* https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf (Fetched 2022-01-31)

[4] American Cancer Society. Targeted Drug Therapy for Breast Cancer. *American Cancer Society*. https://www.cancer.org/cancer/breast-cancer/treatment/targeted-therapy-for-breast-cancer.html (Fetched 20022-04-25)

[5] Audenaert, E. A., Pattyn, C., Steenackers, G., De Roeck, J., Vandermeulen, D., Claes, P., Statistical Shape Modeling of Skeletal Anatomy for Sex Discrimination: Their Training Size, Sexual Dimorphism, and Asymmetry Frontiers in Bioengineering and Biotechnology, volume 7, 2019 DOI=10.3389/fbioe.2019.00302

[6] Breastcancer org. HER2 Status. *Breastcancer org*. 2022.https://www.breastcancer.org/symptoms/diagnosis/her2. (Fetched 2022-01-31)

[7] Brügmann, A., Eld, M., Lelkaitis, G. et al. Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. Breast Cancer Res Treat 132, 41–49 (2012). https://doi.org/10.1007/s10549-011-1514-2

[8] Bröstcancerförbundet. HER2-positiv bröstcancer. *Bröstcancerförbundet*.2021. https://brostcancerforbundet.se/om-brostcancer/vad-ar-brostcancer/olika-former-av-brostcancer/her2-positiv-brostcancer/ (Fetched 2022-01-31)

[9] Can, A., Stewart, C. V., Roysam B., and Tanenbaum, H. L. "A feature-based technique for joint, linear estimation of high-order image-to-mosaic transformations: mosaicing the curved human retina," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 412-419, March 2002, doi: 10.1109/34.990145.

[10] Chalfoun, J., Majurski, M., Blattner, T. et al. MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization. Sci Rep 7, 4988 (2017). https://doi.org/10.1038/s41598-017-04567-y

[11] Dobson, L., Conway, C., Hanley, A., Johnson, A., Costello, S., O'Grady, A., Connolly, Y., Magee, H., O'Shea, D., Jeffers, M. and Kay, E. (2010), Image analysis as an adjunct to manual HER-2 immunohistochemical review: a diagnostic tool to standardize interpretation. Histopathology, 57: 27-38. https://doi.org/10.1111/j.1365-2559.2010.03577.x

[12] Emmenlauer, M., Ronneberger, O., Ponti, A., Schwarb, P., Griffa, A., Filippi, A., Nitschke, R., Driever, W., Burkhardt, H. XuvTools: free, fast and reliable stitching of large 3D datasets. J Microsc. 2009 Jan;233(1):42-60. doi: 10.1111/j.1365-2818.2008.03094.x. PMID: 19196411

[13] Farahani N, Parwani A, Pantanowitz L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathology and Laboratory Medicine International. 2015;7:23-33 https://doi.org/10.2147/PLMI.S59826

[14] Farka, Z., Mickert, M. J., Mikušová, Z., Hlaváček, A., Bouchalová, P., Xu, W., Bouchal, P., Skládal, P., and Gorris, H. H. Surface design of photon-upconversion nanoparticles for high-contrast immunocytochemistry. Nanoscale 12, 15, 8303-8313, (2020). http://dx.doi.org/10.1039/C9NR10568A

[15] Hamilton, P. W., Bankhead, P., Wang, Y., Hutchinson, R., Kieran D., McArt, D. G., James, J., Salto-Tellez, M., Digital pathology and image analysis in tissue

biomarker research, Methods, Volume 70, Issue 1, 2014, Pages 59-73, ISSN 1046-2023, https://doi.org/10.1016/j.ymeth.2014.06.015.

[16] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205387 Fetched 2022-05-04

[17] Jahromi, A.H. and Taheri M, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," 2017 Artificial Intelligence and Signal Processing Conference (AISP), 2017, pp. 209-212, doi: 10.1109/AISP.2017.8324083.

[18] Miyashita, M., Gonda, K., Tada, H., et al. Quantitative diagnosis of HER2 protein expressing breast cancer by single-particle quantum dot imaging. Cancer Med. 2016;5(10):2813-2824. doi:10.1002/cam4.898. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5083734/. (Fetched 2022-02-01)

[19] Muhammad, K., Khan, N., Anil, V. P., Metin, N. G., Digital pathology and artificial intelligence,The Lancet Oncology,Volume 20, Issue 5,2019, Pages e253-e261, ISSN 1470-2045, https://doi.org/10.1016/S1470-2045(19)30154-8.

[20] Nam, S., Chong, Y., Jung, C. K., Kwak, T. Y., Lee, J. Y., Park, J., Rho, M. J., and Go, H. (2020). Introduction to digital pathology and computer-aided pathology. Journal of pathology and translational medicine, 54(2), 125–134. https://doi.org/10.4132/jptm.2019.12.31

[21] Pirovano, A., Heuberger, H., Berlemont, S., Ladjal, S. and Bloch, I., (2020) , Improving Interpretability for Computer-aided Diagnosis tools on Whole Slide Imaging with Multiple Instance Learning and Gradient-based Explanations. ArXiV, 2009.14001.

[22] https://github.com/qupath/qupath Fetched 2022-02-07

[23] Saha, M., and Chakraborty, C., "Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation," in IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2189-2200, May 2018, doi: 10.1109/TIP.2018.2795742.

[24] Wolff, A.C. et al, Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Arch Pathol Lab Med. 2018 Nov;142(11):1364-1382. doi: 10.5858/arpa.2018-0902-SA. Epub 2018 May 30. PMID: 29846104.

A Visual results and examples

A.1 Qupath cell-detection example

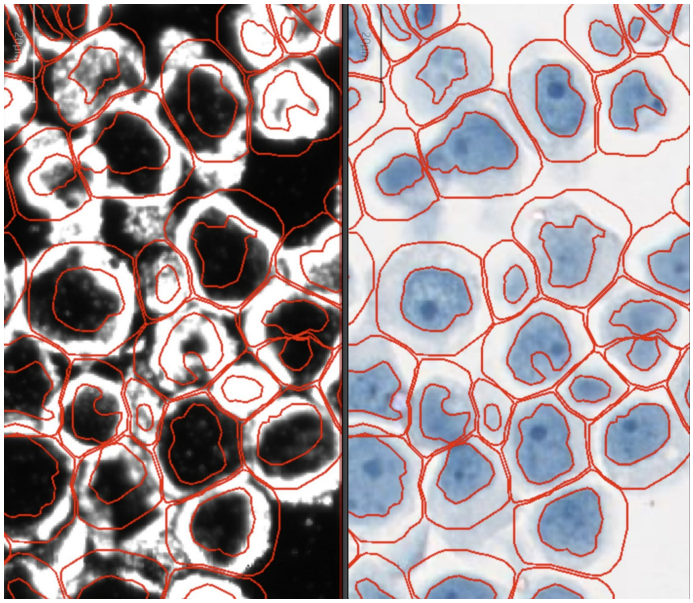


Figure 46: Visual overlook using complete tissue annotation

A.2 Visual overlook using complete tissue annotation

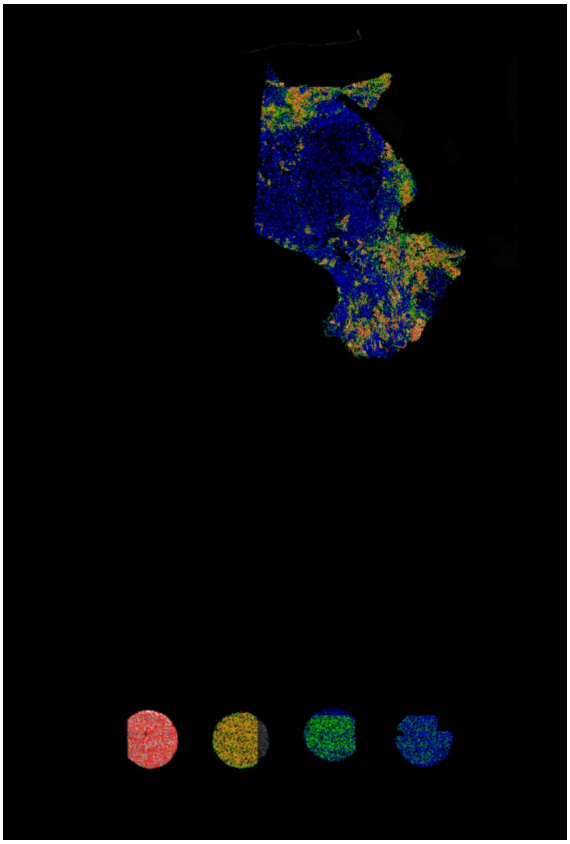


Figure 47: Visual overlook using complete tissue annotation

A.3 Edge effect

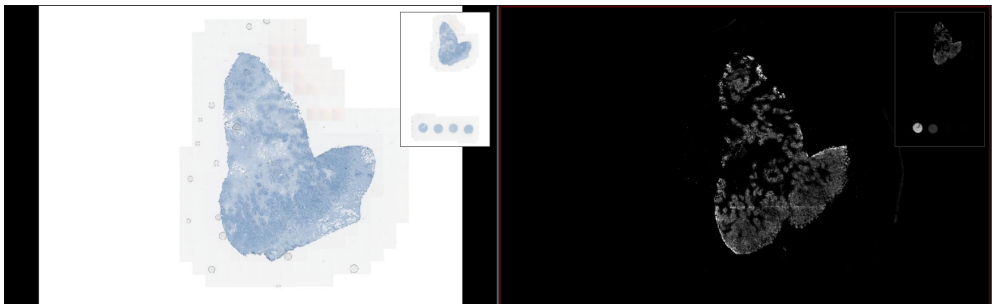
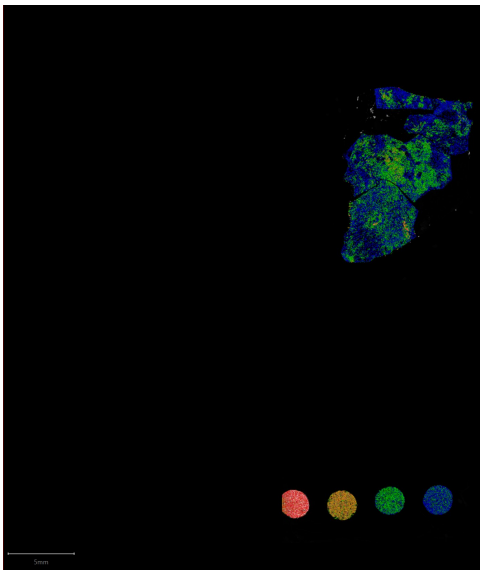


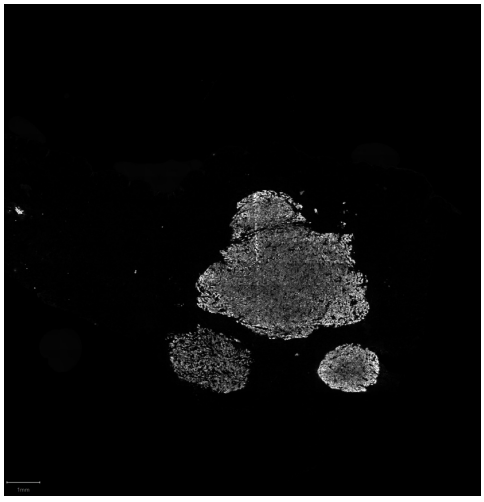
Figure 48: Edge effect example where the edge appear bright in a 2+-sample

A.4 Examples of misclassified images

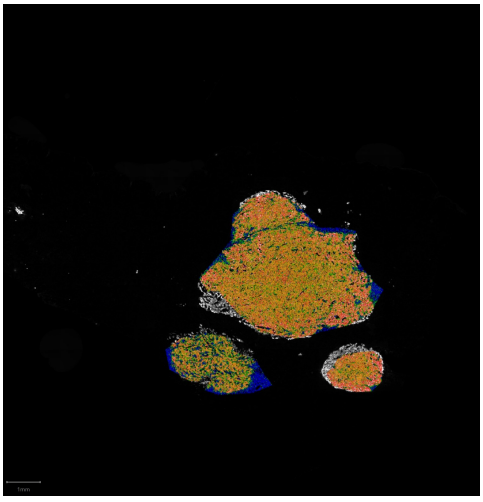
(a) 3+ sample with no classification overlay



(b) Same sample with ipc-overlay



(a) 2+ sample with no classification overlay



(b) Same sample with ipc-overlay

B Qualitative results using all images with complete annotation

B.1 Gaussian classifiers

In Figure 51 the optimal threshold values, using all images as training images, are displayed for the two bayesian models.

Threshold	ratio	number
3+	0.01	6
2+	0.13	230
1+	0.15	170

(a) Threshold values for Bayesian model 512 x 512, only **normalized intensity**

Threshold	ratio	number
3+	0.01	14
2+	0.18	180
1+	0.13	220

(b) Threshold values for Bayesian model 512 x 512, **multidimensional**

Figure 51: Bayesian classifier thresholds on **complete** annotated images using the whole data set

In Figure 52 and 53 confusion matrices for the bayesian models predictions on test and training set combined are displayed. The threshold values: ratio and number, are therefore trained and evaluated on the same images.

	3+	2+	1+	0
3+	9	0	1	0
2+	4	8	0	0
1+	0	0	10	1
0	0	0	7	4

(a) Using number (70.5%)

	3+	2+	1+	0
3+	6	3	1	0
2+	2	10	0	0
1+	0	0	8	3
0	0	0	4	7

(b) Using ratio (70.7%)

Figure 52: Bayesian classifier using only **normalized intensity** for complete annotated images

	3+	2+	1+	0
3+	9	1	0	0
2+	3	9	0	0
1+	0	1	10	0
0	0	0	6	5

(a) Using number (75.0%)

	3+	2+	1+	0
3+	8	1	1	0
2+	3	9	0	0
1+	0	0	9	2
0	0	0	4	7

(b) Using ratio(75.0%)

Figure 53: **Multidimensional** bayesian classifier for complete annotated imaged

B.2 IPC

Threshold	Training set	All images
3+	0.06	0.06
2+	0.04	0.04
1+	0.17	0.18

(a) Cell ratio

Threshold	Training set	All images
3+	14500	14100
2+	8700	8700
1+	59300	69100

(b) Absolute number

Figure 54: Thresholds that optimised confusion matrices for training set and all images respectively

The following are the confusion matrices on all images.

	3+	2+	1+	0
3+	8	1	1	0
2+	2	10	0	0
1+	0	0	10	1
0	0	0	5	6

(a) Cell ratio (77.2 %)

	3+	2+	1+	0
3+	9	1	0	0
2+	3	9	0	0
1+	0	0	8	3
0	0	0	3	8

(b) Absolute number (77.2 %)

Figure 55: Confusion matrices on all images using complete tissue annotation

C Threshold ranges and step size

Table 19: Threshold values for all models. In column 1-3 the threshold values for each threshold are displayed. The range and step size are written as follows (min value, max value, step size). The number of patches or cells will vary dependent on the annotation size, therefore different thresholds needed to be found for ROI-annotations and complete annotations

Model & Annotation method	3+	2+	1+
Bayesian -ROI annotations	(0, 20 ,1)	(0, 1000, 10)	(0,1500, 10)
Bayesian - Complete annotations	(0,100,1)	(0, 1500, 10)	(0, 2500, 10)
Intensity Profile* - ROI annotations	(0, 10000, 100)	(0, 10000, 100)	(0, 10000, 100)
Intensity Profile - Complete annotations	(10000, 20000, 100)	(0, 10000, 100)	(50000, 70000, 100)
Intensity Profile - CPA	(0, 20000, 100)	(0, 20000, 100)	(0, 20000, 100)

*The 1+-threshold search was altered when using all images to 10000 - 20000 with a step size of 100.