Digital Pathology and Computer-aided Diagnostics: a Feasibility Study with Applications to Nasopharyngeal Carcinoma

Sam Forbes, Connah Johnson, Son Le, Steve Tseng

11th June 2018

Abstract

We present an overview of the feasibility of digital pathology and computer aided-diagnostics. Within this, the use of the tumour micro-environment for diagnosis prediction is detailed. Finally, we investigate unsupervised and supervised algorithmic methods on nasopharyngeal carcinoma whole slide images for tumour detection.

Contents

1 Introduction

Digital pathology (DP) is the study of the digitisation and manipulation of histopathological samples with the aim of improving the diagnostic ability of pathologists. Computer-aided diagnostics (CAD) is the development of algorithms to automate aspects of the pathological process. The aim being to provide pathologists with greater insight into the diagnostic information contained within the samples whilst reducing their workload. In this report, we provide an overview of the pathological process and examine the feasibility of the current state of CAD. Additional quantitative information may be derived from samples with a study of the tumour microenvironment (TME) which will be introduced. Finally, the application of unsupervised and supervised machine learning algorithms to nasopharyngeal carcinoma (NPC) cases will provide inferences into the detection of tumour material [1, 2].

2 Digital Pathology

The pathologist's performance is dominated by evaluation speed and diagnosis accuracy, resulting from a high case load with few practicing pathologists. Therefore, increased evaluation speed, accuracy, and improved working conditions are important for histopathology. On a case by case basis, pathologists trained to use the digitised slides saw an increase in slide evaluation speed. One contribution to this could be the fact that the evaluation processes accelerated as the slide does not need to be loaded and unloaded onto a microscope. The use of digitised slides also provide a more ergonomic work condition. Instead of being hunched over a microscope in a darkened room, the pathologist can work in a lighted room in an ergonomic position. This may also assist a pathologist for whom the microscope viewing position may be uncomfortable, such as those suffering from injuries. Remote viewing would also be possible. A pathologist would be able to work away from the lab, reducing the potential for distractions [3].

Currently, up to 60% of slides are digitised, however, due to NHS requirements, screening procedures may not be digitised pending greater research proving the viability of a fully-digitised workflow. The benefits of digitisation become apparent when we look at slide storage. The NHS requires tissue samples to be stored for up to 30 days at which time they are to be destroyed whilst slides are to be kept for 10 years and tissue cassette blocks for 30 years. This produces an ever increasing, large volume of material that must be stored off site for a long period of time. Digitisation would allow for compact storage in a digital form – though this may require an overhaul of NHS computing facilities [4, 5].

3 Computer-aided Diagnostics

Through automation of tedious or highly-quantitative tasks, computer-aided diagnostics (CAD) can assist the pathologist in their daily work. This assistance may be realised through a complete software package that would facilitate the analysis of a WSI, providing numerical measures and guiding tools when required. This would aim to improve clinical diagnosis and pathologists' training and standards. Two recent open-sourced attempts to provide a unified workflow include QuPath [6] and CellProfiler [7]. These provide general image manipulation tools and basic biomarker quantification. While these software packages provide basic CAD tools, more bespoke algorithms may be implemented and are a continuing area of development. Algorithms are produced to analyse different tissue samples with different CAD objectives. The main driving force behind the development of CAD manifests in the form of competitions such as the CAMELYON series of lymph node metastasis challenges or the GlaS challenge for gland segmentation. The main components of produced algorithms include slide pre-processing, thresholding measures and object segmentation. These methods are implemented using both classical machine learning techniques and different neural network architectures [8, 9].

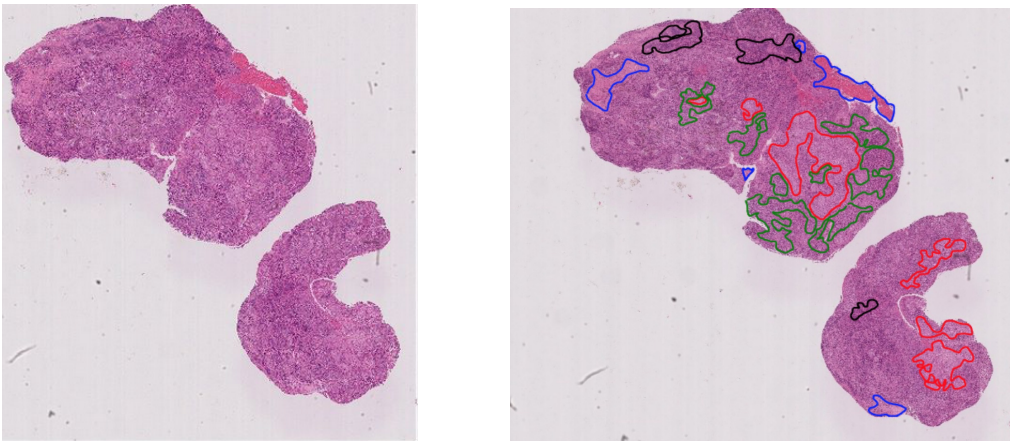


Figure 1: Non-annotated and annotated NPC WSIs. Red: tumour tissue, green: lymphocyte tissue, blue: stroma tissue and black: non-region of interest.

3.1 Pre-processing

As whole slide imaging provides large data files that may suffer from artifacts and varying levels of image quality, these data files must be pre-processed to compensate for imperfections and reduce the memory requirements.

Since the background of a WSI contains no useful diagnostic information (see Figure 1), the first step is to exclude all background, reducing the amount of data that needs to be processed. The background may be removed through considering the patchwise features such as the mean intensity and textural features. A basic segmentation method, such as thresholding, to separate the tissue from the background may also be used.[1].

Staining of WSIs creates a large degree of variability that may be attributed to inconsistencies in staining protocol, individual lab conditions, specimen preparation and the type of scanner used. The staining of an image under consideration may be normalized by forming a histogram of its separate RGB pixel intensities and mapping them to match the histogram of a reference slide. Additionally, one may instead scale the pixel intensities through assuming a Gaussian intensity distribution and scaling via the standard intensity deviation.

One method used to overcome the nonlinear nature of the staining is to map the RGB intensities into optical density space. The map is derived from the Beer-Lambert law which is composed of an exponential map of the decaying light intensity as the light passes through the sample material. The absorbance of each stain and the concentration of each stain attenuates the beam. Therefore the problem of stain variation is cast in terms of the stain and sample properties rather than the non-linear light intensities.

Staining is commonly a combination of multiple stain colours; such as Hematoxylin and Eosin. The separate stain channels may be computationally deconvolved to inspect the distributions of each stain. Here we consider the stained RGB tissue image as the linear product of a stain matrix with the individual stain channels, calculated from the optical density space. The ability to reproduce the stain channels provides greater insights into the biochemical nature of the tissue. This is particularly useful when considering speciality stains such as immunohistochemical Staining (IHC). Methods to do this include; principal component analysis (PCA), sparse nonnegative matrix factorisation (NNMF) and independent component analysis (ICA). PCA produces a set of principal axes with maximal variance to decompose the stain matrix of the RGB channels into individual stain channels. These stain channels can then be analysed individually. NNMF computes two matrices whose elements are non-negative. These matrices are computed such that their product is as close as possible to the input mixed stained image. The rationale behind NNMF is that negative

intensity matrix values are unphysical. Negative values may occur through invoking the orthogonality conditions. The computational expense behind the NNMF routine is dependent on the algorithm used to minimise the difference between the input matrix and the stain channel product. Finally, ICA follows the PCA procedure of determining the stain channel axis however where PCA maximises the variance between the channels, ICA maximises the kurtosis. The benefit of ICA is that the produced stain channels are closer to the true stain channels rather than finding the strictly orthogonal stain channels. This provides benefits regarding the interpretation of the signals at the cost of increased noise when the original stains are of similar colours [10, 11].

3.2 Image Thresholding

The ability to distinguish the tissue sample from the background becomes necessary when considering the removal of background material. The slide digitisation process leads to a high intensity white background with the tissue being of a lower intensity and better inhomogenous RGB channel intensities. Therefore the intensity values may be used to distinguish sample from background. The aim of the thresholding process is to produce a binary mask which identifies the tissue sample from the background slide. This process should also be able to remove image artefacts such as slide barcodes and pathologists annotations. Two commonly used methods to do this are Otsu thresholding and blue ratio thresholding.

Otsu thresholding considers the distribution of pixel intensities over the full whole slide image. The Otsu method is performed on the grayscale intensity map of the image and assumes that the intensities produce a bimodal distribution of the tissue and the background. In this case we may define a threshold value which maximally separates the two distributions. Due to the inhomogeneous colour intensities, the majority of the background pixels will have an intensity greater than the threshold and the majority of the tissue pixels will have an intensity below the threshold. Due to the bimodal nature there would be some error due to the tissue and background distributions overlapping at the threshold, leading to some misclassification error. This error is reduced through minimising the intra-class variance within the two distributions while maximising the inter-class variance. This misclassification risk and the whole slide thresholding approach sees Otsu used when large scale structures are the focus of the analysis. Therefore algorithms highlighting regions of interest and glandular structure tend to favour Otsu thresholding [12].

Blue ratio transform is a method of mapping the RGB intensity values to a single value weighted on the blue channel. The advantage of this over the grayscale transform is based on the Hematoxylin staining of nuclei. One value of diagnostic importance for the pathologists is the mitotic cell count which is based on the number density of nuclei. As nuclei are stained with a blue colour by Hematoxylin, a blue weighted analysis would provide a greater distinction and accuracy of these cells. The blue ratio is often combined with adaptive thresholding techniques which threshold on a patch by patch basis rather than the holistic approach used in Otsu thresholding. Therefore this thresholding technique is suited to nuclei based segmentation analyses [13].

3.3 Region Segmentation

Many of the objectives of CAD algorithms is to segment regions of interest, such as full glandular tissues or individual nuclei. Therefore techniques have been developed to achieve these segmentations. The foundation of segmentation algorithms is a positive point detection method which is then used to locate the related nearby connected points. Three methods of performing segmentation include seeded region growing methods, Bayesian inference of random polygons, and watershed segmentation.

The region growing method searches the image space for features of the desired target cell or gland that are absent in the surrounding material. An example in the literature used an algorithm to search for large vacant inner regions of cells surrounded by a chain of smaller cells. The vacancy forms the seed of the method with the chain of cells being a boundary condition. Other boundary conditions may be a rapid change in intensity of the structure signifying a cell membrane. The algorithm dilates the positively defined structure to form the union of pixels that compose the structure. Starting from the seed; connected pixels are successively

added to the growing structure up to the boundary where growing is terminated. The number of iterations for the structure to grow may be used to define a growing age to prune ill bounded regions or as a termination condition [14].

The Bayesian inference of random polygons looks to surround the structure of interest with a polygon. The optimal polygon is found through considering the bayesian inference of a space of possible n vertex polygons and a prior probability that depends on normally distributed lengths and corner angles of the lines composing the polygons. The locations of these polygons is seeded through the maxima of a probability map of the structure of interest, with vertices placed at the individual nuclei points. The maximum a posteriori polygons are found using a Markov chain Monte Carlo method allowing reversible jumps and optimised through the Metropolis-Hastings numerical method. The results of the algorithm are the possible structure polygons. False positives are reduced through an Occams razor like heuristic removing over complex number of vertices and pixelwise large polygons. Benefits of the random polygon method are an inherent parallelism with each polygon being determined independently of each other. There is also a robustness to pixel noise within the structures themselves as they consider the boundary only [15].

Finally the watershed method is used to distinguish overlapped structures. A local maxima finding algorithm or other seed determining methodology is applied to define the centers of the structures. The distance of the seeds to the nearest background pixel is computed and compared to a predefined value for the minimum structure size. Nearby seed distances smaller than the minimum threshold are merged to denote one structure, reducing the fragmentation. The distance against intensity space produces a surface where the local maxima and minima highlight the structures. Therefore the seeds are then grown until the edges of the boundaries are reached analogous to the geophysical phenomena of rising water filling a valley. Therefore the watershed method is robust to overlapping regions and disconnected structure segments. The output contours may be smoothed using filters and overfitting reduced by using morphological metric thresholds [16].

3.4 Deep Learning Architectures

This CAD study is focused on deep learning techniques. Deep learning algorithms are based on a variety of foundational architectures which are tasked with classifying patches of the WSI. To reach this aim the networks take a convolutional nature to process images. Therefore, to fully utilise these architectures a large annotated image dataset is required. As such a large dataset is rare in the digital pathology field the network are usually pre-trained on a large computer vision database, such as ImageNet. The pathological dataset is then used to perform fine tuning on the final layers of the otherwise trained CNNs.

A variety of CNNs were utilised in the surveyed work. A selection of these were trained on a series of NPC image patches with the aim of classifying tumour and non-tumour regions. These networks were pretrained on the ImageNet dataset [17] before fine tuning the final 2 and final 3 layers separately. Using a further testing set from the same NPC slide the accuracy and F1 metrics were calculated.

	Accuracy	F1 scores	Model
0	0.69373	0.703125	Xception_2
1	0.70072	0.703125	Xception_3
2	0.77895	0.828125	VGG16_2
3	0.71955	0.765625	VGG16_3
4	0.76789	0.733333	VGG19_2
5	0.77232	0.718750	VGG19_3
6	0.36900	0.406250	ResNet50_2
7	0.36900	0.281250	ResNet50_3
8	0.74908	0.812500	InceptionV3_2
9	0.36900	0.333333	InceptionV3_3
10	0.73949	0.703125	InceptionResNetV2_2
11	0.71142	0.734375	InceptionResNetV2_3

Figure 2: Table showing the accuracy and F1 scores for each of the trained models based on the selection of architectures. The model suffix 2 or 3 indices the number of fine tuned layers.

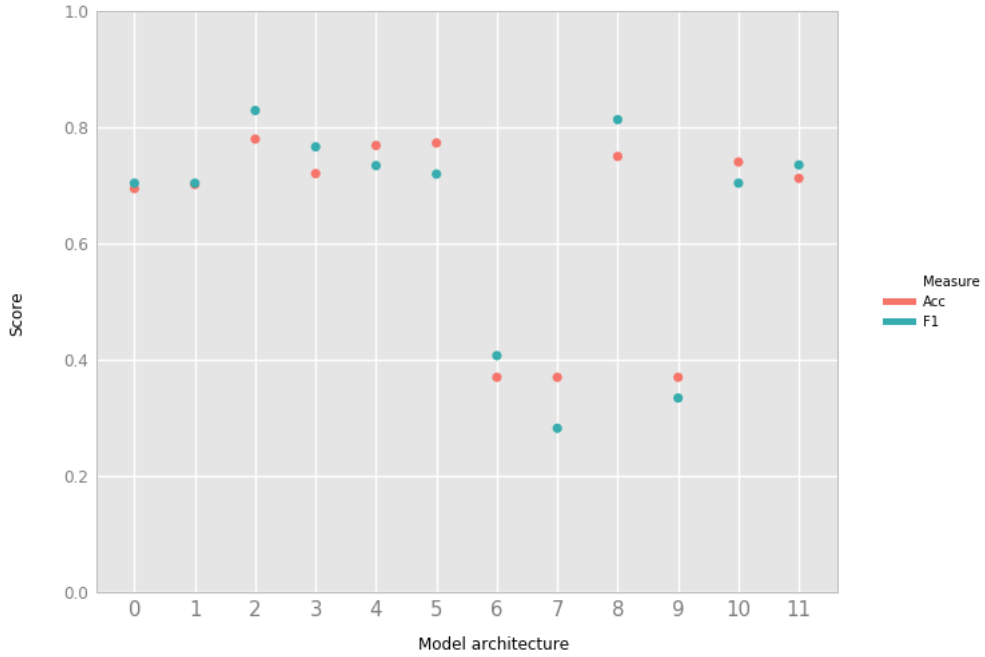


Figure 3: Scatter plot showing the accuracy and F1 scores for the model architectures defined in Figure 1. Model 2 gives the highest accuracy and F1 scores corresponding to the VGG16 model fine tuning the final 2 layers. In general the 3 layer models have decreased accuracy and F1 scores suggesting that the fine tuning dataset was insufficient in size for training the higher layers.

The results, given in Figure 2 and shown in Figure 3, suggest that the best performing architecture, of the selection tested and on the training/testing data sets provided, was the VGG16 model with the final two layers being re-trained on the dataset.

3.5 CAD Algorithm Survey

The aforementioned surveyed works were selected to build a picture of the state-of-the art in the CAD field. The field suffers from a lack of standardisation as a whole. Different groups tend to use their own datasets sourced from few clinical centres. Reporting measures are also not standardised, furthering this disparity. Results that may be compared face to face tend to be clustered in competition results working towards the same objectives. These competitions also provide larger annotated datasets increasing the quality of the results. The presented surveyed papers span the 2015-2017 spanning different objectives and different methodologies [18, 19, 20, 15, 21, 22, 24, 25, 16, 13, 26, 27].

Author	Year	Objective	Tissue	Pre.processing	Method
D. Wang	2016	Detect metastatic tissue	Breast, sentinel lymph nodes	Remove background, RGB to HSV, Otsu	Patch based deep learning, googlenet, then extract 28 features with random forest classifier
G. Litjens	2016	Pixelwise cancer likelihood map	Prostate cancer	Data augmentation and boosting, simple thresholding on OD space	Patch extraction train CNNs, convert to fully convolutional network for runtime
G. Litjens	2016	Pixelwise cancer likelihood map	Breast, sentinel lymph nodes	Data augmentation and boosting, simple thresholding on OD space	Patch extraction train CNNs, convert to fully convolutional network for runtime
Y. Zhang	2017	Gland segmentation of unannotated images	Colon cancer		Two networks; segmenation follows DCAN and evaluation follows VGG16
Y. Zhang	2017	Fungus segmentation of unannotated images	Ant body tissue		
K. Sirinukunwattana	2015	Gland segmentation	Colon cancer	Normlaise stain intensities, crop background slide	Bayesian inferences of random polygons
N. Kumar	2017	Generalized nuclear segmentation	Breast, liver, kidney, prostate, bladder, colon, stomach	Colour normalisation	Nuclei detection classifying in/out of cells using CNN, isotropic region growing
N. Kumar	2017	Generalized nuclear segmentation	Breast, liver, kidney, prostate, bladder, colon, stomach	Colour normalisation	Nuclei detection classifying in/out/boundary of cells using CNN, isotropic region growing
Y. Liu	2017	Detect metastatic tissue		Image augmentation	Tumor probability map using patch based inference in Inception V3 CNN
K. Paeng	2016	Tumour proliferation scoring		WSI resizing, otsu thresholding, stain normalisation	Select top 30 patches of highest cell density, deep cnn based on ResNet, encode into feature vectort then SVM to predict score
T. Qaiser	2017	Her2 scoring	Breast cancer		
N. Alsubale	2017	Nulcei detection	Colon cancer	ICA in wavelet domain producing H channel	Spatially constrained CNN detect centroids
R. Paulik	2017	Nuclei detection	Breast cancer	Extract nuclei channel from background, adaptive thresholding, box filtering	Local maxima finding cell centres, watershed segmentation, seed growing
D.Romo-Bucheli	2017	Mitotic cell quantification		Adaptive threshold on blue ratio	CNN to assign probability patch contains mitotic event
A. Lindberg	2017	Proliferation heat mapping	Breast cancer	Sliding wwindow to compute pixel probabilites	Threshold connected areas with maximum value, dilate seed grow
M. Valkonen	2017	Detect metastatic tissue	Breast, sentinel lymph nodes		

Figure 4: Table showing the methodologies employed in the literature. The variety and scope of the objectives and tissue groups may be seen on inspection.

Author	Year	Dice.score	PPV	Recall	AUC	Number.of.scanners	Number.of.Images	Number.of.cases	Number.of.centres	Train.test.split
D. Wang	2016	NA	0.98	NA	0.93	2	400	NA	2	270/130
G. Litjens	2016	NA	NA	NA	0.99	2	225	225	1	150/75
G. Litjens	2016	NA	0.93	NA	0.88	2	271	NA	1	131/140
Y. Zhang	2017	0.92	NA	NA	NA	NA	245	NA	NA	185/60
Y. Zhang	2017	NA	0.94	0.93	NA	NA	NA	NA	1	40/4
K. Sirinukunwattana	2015	0.87	NA	NA	NA	2	146	NA	2	72/74
N. Kumar	2017	0.69	0.78	NA	NA	NA	30	30	18	12/14
N. Kumar	2017	0.76	0.92	NA	NA	NA	30	30	18	12/14
Y. Liu	2017	NA	NA	NA	0.98	2	400	NA	2	270/130
K. Paeng	2016	NA	NA	NA	NA	NA	73	73	3	656 patches/34 patches
T. Qaiser	2017	NA	NA	NA	NA	NA	172	86	NA	NA
N. Alsubale	2017	NA	0.81	0.42	NA	2	27	NA	2	7/20
R. Paulik	2017	NA	0.90	0.88	NA	1	30	25	1	NA
D.Romo-Bucheli	2017	NA	0.47	0.78	NA	2	741	NA	2	NA
A. Lindberg	2017	NA	NA	NA	NA	1	56	NA	NA	NA
M. Valkonen	2017	NA	NA	NA	0.98	NA	NA	NA	NA	NA

Figure 5: Table collating the algorithm scores and data set information as reported in their respective papers. The high number of missing values stems from the non-standardisation between reporting measures. The high disparity in the data set size and quality, number of centres and number of cases, may be seen. A high quality dataset would have greater values for all the data numbers however this is rare in the literature.

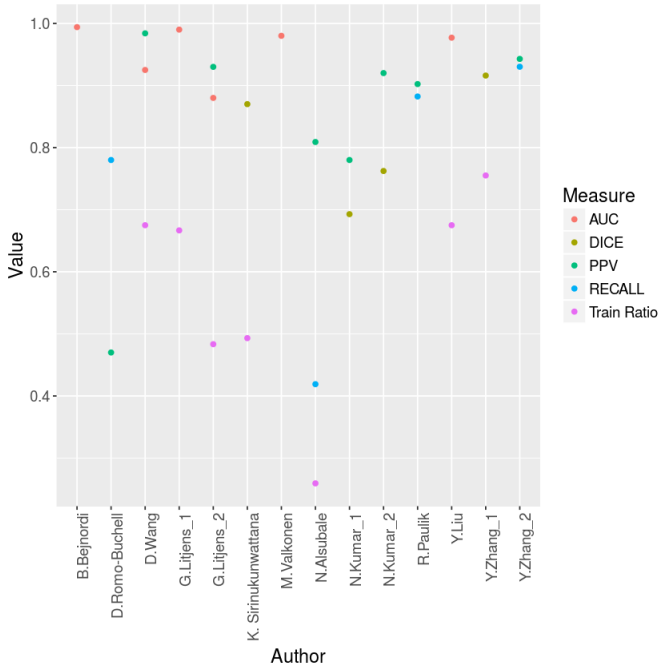


Figure 6: Scatter plot graphically showing the algorithm metrics presented in Table 2. The ratio of training to testing image data is presented via the Train Ratio. Due to the sporadic nature of the published metrics comparisons may only be drawn in clusters using each metric separately.

The results of the algorithmic feasibility study portray an underdeveloped field. The lack of standardisation of reported metric choices and datasets leads to a weakening of conclusions that may be drawn from field wide studies. The high variance in the results may be indicative of large changes in methodology required when considering different tissue types. One may see a lower accuracy in the algorithms attempting to generalise to multiple tissue types. Consequently, the specialised high accuracy algorithms may be considered as overfitting from the most general case. When we consider the clinical deployment of CAD algorithms having a large suite of specialised algorithms would lead to large overheads in validation, as each algorithm will have to reach the gold standard in turn. Therefore these general approaches may find it easier to garner support from the pathology community and lead to a more feasible route for CAD. This conclusion may be strengthened when considering the views of the pathology community.

4 Pathologist Survey: Use of Machine Learning in Digital Pathology

Computer assisted diagnostic is by nature an industry aligned field where the aim is to get algorithms and software ready for clinical deployment. Therefore to focus solely on the techniques may produce feasibility conclusion with a heavy algorithmic bias. Therefore the views of pathologists should provide a holistic view of the field and a direction for progress to take. Therefore, in line with past digital pathology focused surveys, a Computer-aided diagnostics survey was distributed to gather these views. The survey consisted of 9 questions varying between multiple choice and free text types. The survey was written using Survey Monkey and distributed via a leading pathology online publication website and through pathologists social media. The aim of this was to aim the survey to a group of pathologists that have not necessarily been exposed to CAD before, controlling potential biases. The survey received 46 responses over a 72 hour period with a third self declared Academic Clinician, 28.6% Clinicians, 23.8% Academics with the rest being trainees, student or other denominations [28, 29].

Number	Question
Q1	What is your background and level in Pathology?
Q2	How often do you use digital slides in your work?
Q3	What does digital pathology and CAD mean to you?
Q4	What aspects, if any, of CAD do you currently employ?
Q5	Do you feel that there are daily tasks that could be improved with the assistance of software?
Q6	Do you feel that the introduction of CAD could improve working conditions in histopathology?
Q7	How well does CAD meet any safety concerns with the automated analysis of clinical slides?
Q8	Do you feel CAD is ready for widespread adoption in histopathology?
Q9	Do you think machine learning and artificial intelligence can benefit the histopathology field?

Figure 7: Table showing the questions presented to pathologists in a Survey Monkey survey. Q1 was used to group the respondents into backgrounds for the analyses, Q2 was a free text question to gather personal views, and the remaining questions were multiple choice.

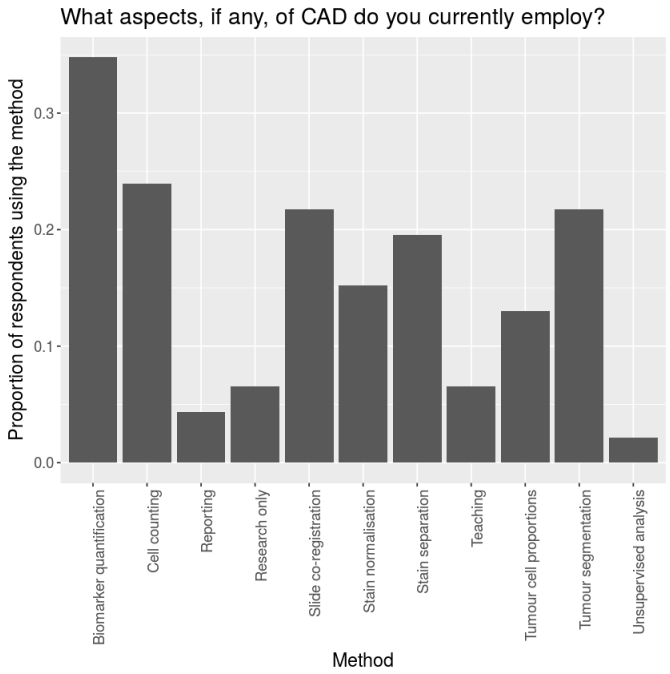


Figure 8: Bar plot showing the respondents currently employed CAD objectives. The survey encouraged a declaration of all tools used and then these counts were recorded as a fraction of the total number of respondents, 46. The most used tools were for performing tasks seen as tedious and repetitive leaving the pathologist more time to consider more complex metrics.

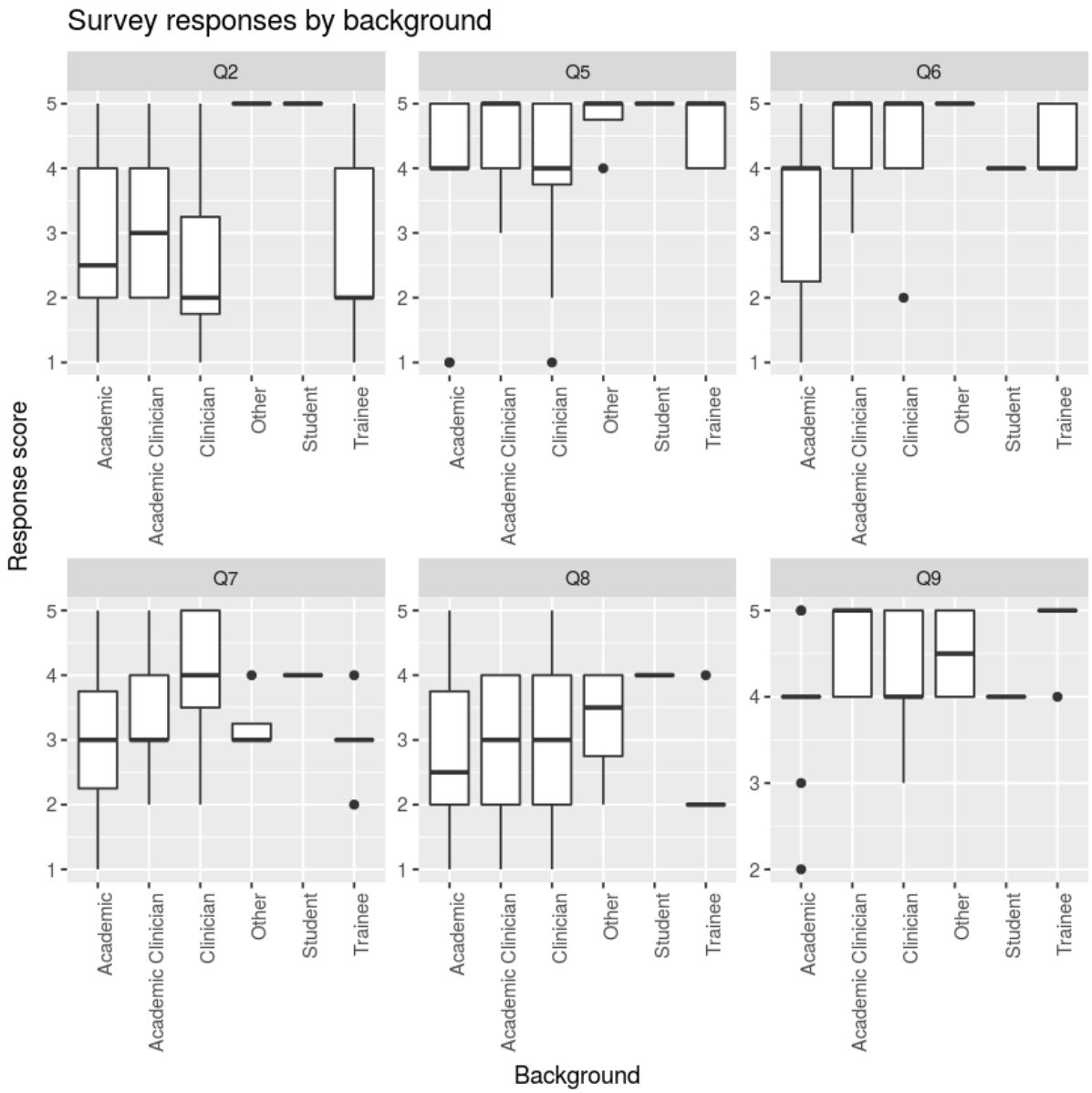


Figure 9: Multi-facetted box plot showing the results of the multiple choice questions. The survey participants were asked to score their response 1 for strongly negative, 3 for neutral, and 5 for strongly positive. The thick black line representing the median score may be isolated in instances with low response numbers or very polar opinions. A key for the question numbers may be found in Figure 6.

The free text questions provided the survey participants an opportunity to state their opinion unconstrained. The overall opinion of the pathologists surveyed is that CAD promises to provide a useful selection of tools to increase efficiency and accuracy which would improve the diagnostic workflow holistically. This is seen to be through increasing the information provided to the pathologist while enabling greater quality controls and eliciting further opinions. This is corroborated with an overall positive opinion around the technology itself and with the opinion that machine learning can benefit pathology being almost universally agreed. However this is against the backdrop that CAD isnt seen to be ready for widespread adoption in the current generation of technologies. This is furthered by the tentative guidance regarding CAD by the Royal Pathological Society.

Optimistic conclusions that may be drawn from the survey may include the positive reception to CAD

and machine learning in general. A majority of respondents presented an enthusiastic opinion of CAD with promising advancements to pathology. This may suggest that the ability to acquire larger datasets of annotated WSIs for algorithm training may become easier as pathologists may be more willing to collaborate in the immediate future.

5 Tumour Microenvironment

Numerous studies have shown that humans can live decades with malignancies within their bodies without any manifestations of overt cancer [30]. Life-threatening tumours develop four key traits: the capacity to breakdown the extracellular matrix (ECM), the facility to migrate, the fitness to survive in the bloodstream, and the ability to establish itself in different tissue, i.e., metastasize [31]. Through junctions, receptors, hormones and other soluble factors, cells communicate with each other and components of the ECM to maintain the cellular microenvironment and proper biological processes, including immune suppression of tumour growth [30]. Degradation of cellular signaling, caused by various factors including exposure to carcinogens and inflammation, inhibits the detection of tumours, allowing cancer cells to flourish and reshape the environment in its favour – even recruiting host cells to support its progression and propagation [32]. From studying the cellular environment of the tumour, the tumour microenvironment (TME), we will be able to ascertain the mechanisms that allow tumour cells to overcome the immune systems defenses and manifest as cancer.

5.1 Components of the Tumour Microenvironment

The TME materializes through tumour-host interactions with the tumour acting almost like an organ as it dictates molecular and cellular activity around the encompassing tissue [33]. It is a complex structure consisting of a variety of cell types including malignant cells, lymphocytes, endothelial cells, myeloid cells, natural killer cells, fibroblasts, pericytes, adipocytes and stem cells [34]. The TME also includes the vascular structures of blood and lymphatic vessels and the ECM [34]. The intercellular interactions of the TME are facilitated by various chemicals and enzymes such as cytokines, chemokines, growth factors and matrix-remodelling enzymes [34].

Through a process called tumour immune surveillance, $CD8^+$ or cytotoxic T lymphocytes (CTL), with the support of $CD4^+$ T helper 1 (TH1) can identify tumour cells and annihilate them [35]. These tumour-infiltrating lymphocytes (TIL) account for a large part of the TME and when in abundance, is often strongly associated with a good prognosis [36]. However, there is also conflicting evidence of large quantities of TILs in the TME supporting a negative prognosis [36]. Oncologists suspect that this may be due to cancer cells suppressing the tumour-killing abilities of CTLs and expect that further study of the TME will elucidate on this contradiction [33].

5.2 Tumour Microenvironment of Nasopharyngeal Carcinoma

Undifferentiated NPC is a highly metastatic cancer where by the time of diagnosis, many patients already have lymph node metastases [37]. This may be explained by the distinct and heavily-heterogeneous TME of EBV-associated NPC where heavy infiltration of TILs, accounting for 50% of the tumour mass, coexist alongside EBV-infected NPC cells [37, 38]. Moreover, cytokines, chemokines and exosomes secreted and released by EBV-infected cancer cells are present in large quantities in the TME [37]. Through these signaling factors, oncologists believe that NPC cells communicate with stromal cells to overcome and evade immune surveillance, resulting in metastases [37]. Furthermore, viral products such as miR-BARTs, a microRNA, and BARF1, a protein, are also found in abundance in the TME of NPC [37, 38]. It is believed that these viral products adjust the gene expression of host cells for negative regulation whilst promoting NPC cell immune evasion, proliferation, cell apoptosis and metabolism [38, 39].

5.3 The Role of Digital Pathology

The spatial context of the TME is pivotal to understanding how tumour-host interactions promote cancer growth [40]. With the availability of WSIs of tumour sections, computer-assisted analysis can now be applied to extract information such as the heterogeneity in specific parts of the tumour, measure distances between TME components or phenotype vast quantity of cells [40, 41].

For example, through WSIs of gastric adenocarcinoma, Feichtenbeiner et al. in [42] sought to shed light on conflicting data regarding the benefit of large quantities of TILs within the TME by counting the number of $CD8^+$ and $FoxP3^+$ TILs and measuring their cell-to-cell distances [42]. Feichtenbeiner et al. hypothesized that a positive prognosis due to the antitumor effects of CTLs was dependent on the spatial distribution of TILs [42]. Their results showed that patients with $CD8^+$ - to - $FoxP3^+$ distances between 30 - 110μm had a ten-year survival rate of 88.7% compared to 55.4% when the distances were outside this range [42]. Additionally, they found that high quantities of TILs only improved survival rates when the TILs were within the epithelial compartment of the TME [42]. Hence, the co-localization of TILs was necessary to their biological effect, see Figure 10.

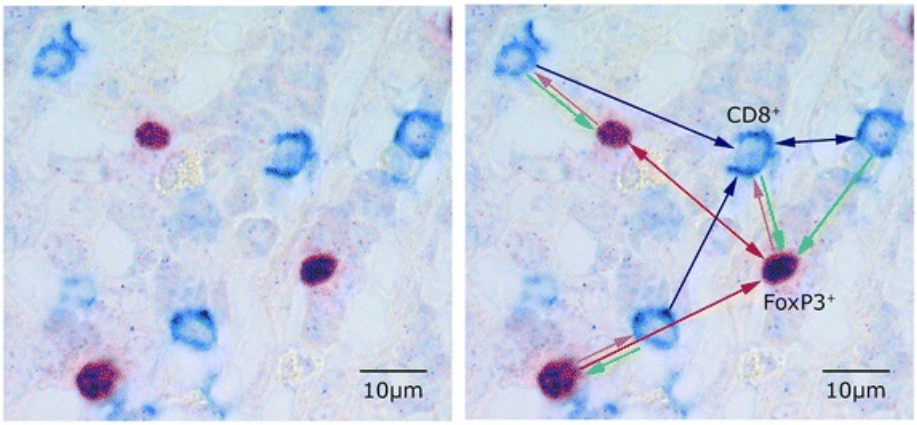


Figure 10: Left: A patch from a whole slide image of gastric adenocarcinoma that has been immunohistochemically stained to show the $CD8^+$ (in blue) and $FoxP3^+$ (in purple) tumour-infiltrating lymphocytes. Right: Distances to the nearest cell indicated by red arrows $FoxP3^+$ - to - $FoxP3^+$, blue arrows $CD8^+$ - to - $CD8^+$, orange arrows $FoxP3^+$ - to - $CD8^+$, and green arrows $CD8^+$ - to - $FoxP3^+$ [42].

As demonstrated in [42], digitised histological images allow us to discover information regarding the behavior of cancer cells and their effect on their surroundings that was once otherwise unknown. Moreover, the application of mathematical and statistical models already performed on histological data such as, using the Morisita-Horn index to determine whether a predator-prey relationship exists between cancer cells and immune cells and the creation of a hybrid discrete-continuous agent based model to simulate the spatial disorganization of the TME, shows that histological data is a mine of biological information ripe for computational analyses [43, 44]. With DP still in its infancy, the development of objective and standardized methods, eliminating subjective bias, to analyze TME components will allow us to transform histological data into biological markers and new therapies.

Additionally, traditional forms of treatment are focused on the removal of proliferating tumours through radiotherapy, chemotherapy and surgical resection, and often times are toxic, ineffective in late-term cancer and have a high rate of recurrence [45, 47]. Digital pathology could be utilized to observe the changes in the quantities of tumour-suppressing and tumour-promoting components in the TME, thereby monitoring the effectiveness of these treatments [46]. Perhaps more significantly, the capacity of DP to elucidate the spatial context of the TME, permitting for greater understanding of tumour-host interactions, could be crucial for developing treatments that exploit a patients own immune system to return the TME back to its homeostatic state [47]. One such treatment, immunotherapy, has proven to be effective during late-term cancer and

would be invaluable for the treatment of NPC as NPC metastases is common and is the predominant cause of treatment failure [48, 37]. Accordingly, there is much potential for DP to enhance our knowledge of tumour-host interactions and facilitate discovery of new cancer treatments.

6 Data

For this project, we had access to five hematoxylin and eosin (H&E) stained WSIs of NPC tissue. Two of these had pathologist annotations with the first annotated for tumour cells, stroma, lymphocytes and non-region of interest whilst the second for tumour cells and non-region of interest (see Figure 11).

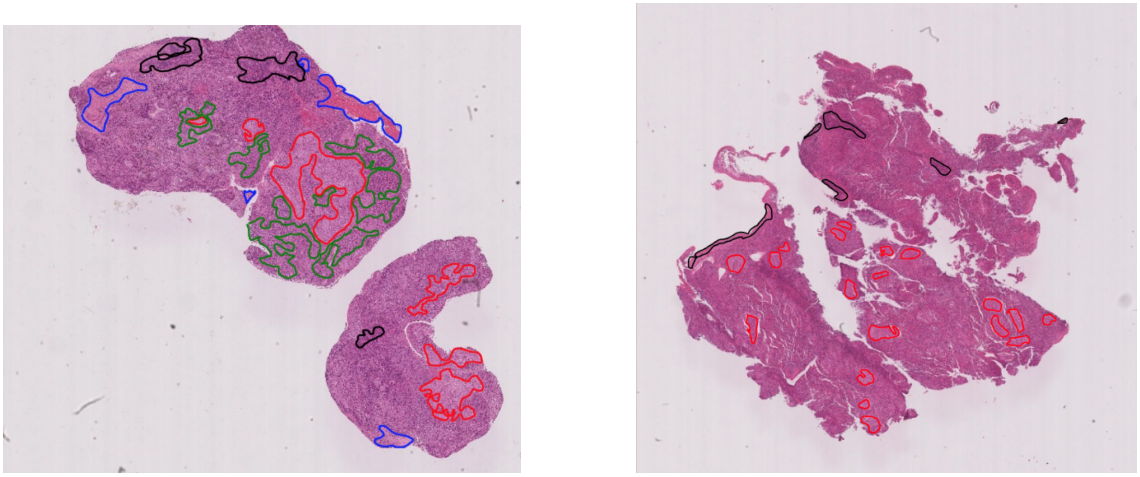


Figure 11: Nasopharyngeal carcinoma whole slide images from two patients with annotations for tumour tissue (red), lymphocytes (green), stroma (blue) and non-region of interest (black).

Our WSIs had 6 magnification levels with the highest magnification, level 0, having an image size of $106259 \times 306939 \times 3$, where the 3 refers to the red, green, blue (RGB) colour intensities. For analyses, the WSIs were sectioned into patches (or tiles) of $256 \times 256 \times 3$.

Additional data from the Computational Precision Medicine (CPM) Nuclei Segmentation Challenge at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017 conference for nuclei segmentation was also utilised. The dataset consisted of 32 training and 32 testing image tiles, extracted from multi-tissue H&E stained histology slides, with an equal representation of glioblastoma multiforme (GBM), low-grade glioma (LGG), head and neck squamous cell carcinoma (HNSCC) and non-small cell lung cancer (NSCLC). Segmentation ground truth for all nuclei was provided in the form of a binary mask.

7 Methods

7.1 Features

Suitable features must be extracted from the images to obtain an accurate classification. Object-level and textural types of features are often used in analyses of WSIs [3, 56].

7.1.1 Object-level

Object-level or morphological features describe the size, shape, eccentricity and other geometric traits of specific objects in an image. Accordingly, these features are useful in detecting particular objects in an image such as nuclei [53].

7.1.2 Textural

Textural features describe how fine, coarse or irregular patterns are in an image. As the texture of an image is related to its tone through the spatial distribution of grey-scale pixel values, textural features may be extracted by converting an image into grey-scale [56]. Extracting these textural features produced a 74-length vector.

Central moments are the different statistical moments, such as the mean, variance, skewness and kurtosis, of the pixel's intensities. The invariance of central moments under translations of the image space makes them especially useful [56].

Local Binary Patterns (LBP) capture the local spatial structure of an image by applying an activation kernel to each pixel with its surrounding 8 neighbors. The value of the pixel is taken as the binary threshold compared to the value of the 8 neighbors. The procedure results in a binary array of length 8 for each pixel.

The grey-level co-occurrence matrix (GLCM) is the distribution of co-occurring intensity values at a given pixel offset. GLCM calculates how often a pixel with grey-level value i occurs adjacent to pixels with the value j. Using GLCM, various statistics such as intensity contrast and correlation, energy and pixel homogeneity can be derived.

Perceptual features are based on the contrast, directionality, line-likeliness, contrast and roughness of the image.

7.2 Clustering Algorithms

7.2.1 Density-based Spatial Clustering of Applications with Noise

Density-based spatial clustering of applications with noise (DBScan) assigns points of the feature space into particular clusters or designates them as noise. The cluster assignments are based on the local density of each point using two parameters: $minPts$ and ε. If a point has at least $minPts$ other points within a radius of ε then it is designated a *core point*. Formally, given a set $P = \{p_1, p_2, \ldots, p_N\}$ of N points in M dimensional feature space, p is a *core point* if there exists a minimum number $minPts$ of points $p_j \in P \setminus p$ such that $\|p - p_j\| < \varepsilon$ where $\|x\|$ is the M dimensional Euclidean distance. Two core points p_i and p_j with $i \neq j$ belong to the same cluster c if $\|p_i - p_j\| < \varepsilon$. These two core points are then specifically core points of c: $p_i, p_j \in core(c)$. A point that is not a core point but lies within a distance of ε from a core point is called *border point* and belongs to the cluster of that core point. All remaining points, neither core or border, are classified as *noise*.

The algorithm is perhaps more easily illustrated as a diagram: 8 points in Figure 12 with $minPts = 3$ and $\varepsilon = 2.5$cm are clustered with DBScan. One cluster is formed containing 4 core points and 3 border points leaving one remaining noise point.

A problem with using DBScan is selection of the two parameters $minPts$ and ε which one usually chooses in a trial and error approach. However, this can be costly with large amounts of data. Another disadvantage is that DBScan may classify a high percentage of points as noise. However DBScan clusters the points into k clusters without the need to predetermine k.

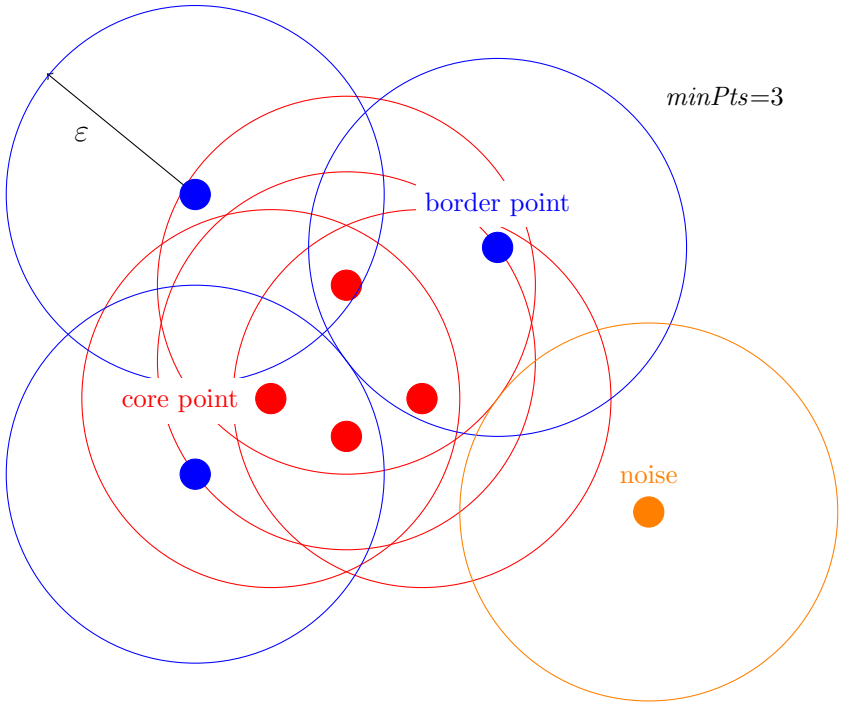


Figure 12: Illustration of DBScan clustering 8 points with $minPoints = 3$ and $\varepsilon = 2.5$cm.

7.2.2 k-means Clustering

k-means clustering algorithm is an iterative procedure to partition the feature space points into k clusters by minimising the within cluster variance. The steps of the algorithm can be defined as follows:

1. With k specified, k means $\{m_1^0, m_2^0, \ldots, m_k^0\}$ are randomly initialised and each point in the feature space $\{p_1, p_2, \ldots, p_N\}$ is assigned to the closest m_i^0 to form k clusters $\{C_1^0, C_2^0, \ldots, C_k^0\}$.
2. The mean of each cluster is computed and these form the new means.
3. Step 1 is carried out with these new k means. The process terminates when there are no further changes of cluster assignments of the points.

The update rule to find the new means is defined (the centre of mass of the cluster):

$$m_i^{s+1} = \frac{1}{|C_i^s|} \sum_{p_j \in C_i^s} p_j.$$

If two means are equally close to a point, the point is assigned arbitrarily. If no points are assigned to a cluster then it is removed.

To determine the optimum number of classes, k, a heuristic method is to locate the 'elbow' of the graph of the 'number of clusters' versus the 'corresponding total within cluster variance'. Within cluster variance usually decreases as cluster size increases. When the decrease begins to level off, this indicates that the optimum number of clusters are starting to be partitioned into further clusters. Thus, the optimum number of clusters, k, corresponds to the 'elbow' of the graph (see Figure 13 (b)).

7.3 Classification

7.3.1 Convolutional Neural Networks

A convolutional neural network (CNN) is a deep, feed-forward artificial neural network (ANN) with at least one layer using convolution in place of matrix multiplication. Unlike a regular ANN, a CNN has 3-dimensional layers: width, height and depth. A filter performs the convolutions on the input, whose sum is transfered to the feature map. The feature maps become the inputs to the activation functions. A pooling layer for reducing dimensionality is common between CNN layers. This architecture makes CNN prime for feature extraction and particularly good for applications involving images [56].

The U-net model is a CNN that relies on data augmentation to reduce training data requirements [50]. The crucial property of the model is that it balances between finding the localization of high-level features versus capturing context. These properties make the U-net particularly good for biomedical image segmentation [50].

The vgg-16 CNN is a very deep configuration consisting of 16 layers and made well known by it's winning performance on the ImageNet Challenge 2014 [51].

7.3.2 Random Forest

The random forest algorithm is an ensemble machine learning approach using bagging or bootstrap aggregation. The Random Forest algorithm constructs an ensemble of decision trees and takes the mode (classification) or the average (regression) of those outputs as a result. In general, decision trees that are not pruned tend to overfit the data which results in a low bias but a high variance. Taking the average of many models reduces the variance without increasing the bias. This increases the performance of the ensemble method significantly. The approach works well as long as the individual trees are not strongly correlated. Strong correlation is avoided by using different subsamples with replacement of the overall training set for each decision tree.

7.3.3 Support Vector Machine

The support vector machine (SVM) is a binary classification model which constructs a linear discriminant decision boundary. The algorithm maximizes the margin which is the perpendicular distance between the decision boundary and the closest points on each side. Intuitively, the SVM algorithm finds the hyperplane which maximizes the distance to the nearest data point on each side. Therefore the decision boundary only depends on these nearest points which makes this model very sparse. The result is that the prediction time is rapid as only a small number of observations are retained and utilized for predictions.

8 Unsupervised Learning: Clustering of Nuclei

Upon the start of our research project, we had not yet received annotated NPC WSIs from the pathology lab. Therefore, we decided to utilise an unsupervised learning method to extract nucleic features of NPC cells and utilise them to cluster the nuclei into classes. As the appearance of nuclei in histological images are critical indicators of cancer, determining the differentiation between the nuclei of various cells in NPC tissue would be advantageous during diagnoses [53, 18].

8.1 Nucleic Features

The nuclei of NPC cells are hyperchromatic, i.e., stain much darker than the nuclei of nonmalignant cells, have an irregular shape, have high nuclear-to-cytoplasmic ratio and contain prominent nucleoli (see Figure 14) [52]. Since our goal was to detect and classify nuclei, we chose to extract the following nine object-level features for each nuclei:

1. x-coordinate of centre of mass

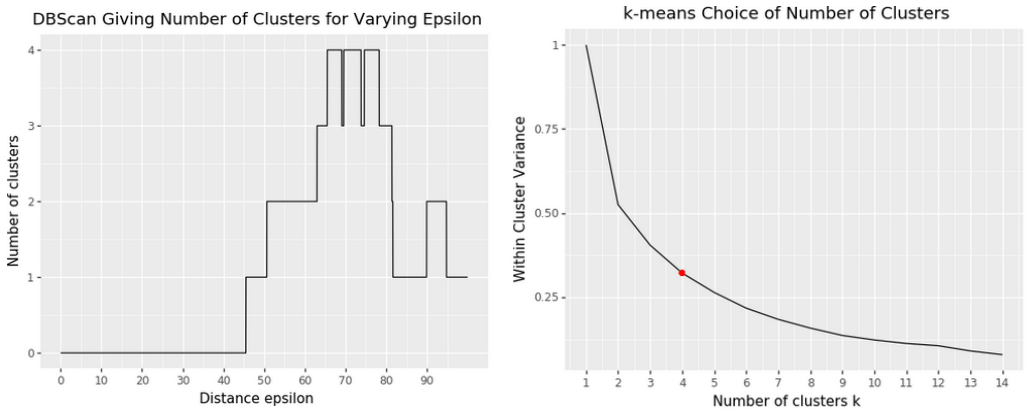
2. y-coordinate of centre of mass
3. Pixel-wise size
4. Eccentricity (a measure of the deviation away from a circle)
5. Mean intensity of greyscale background
6. Mean intensity of greyscale nuclei
7. Standard deviation of intensity of greyscale nuclei
8. Smallest greyscale intensity value in nuclei
9. Highest greyscale intensity value in nuclei.

8.2 Feature Extraction

To identify nuclei, the U-Net CNN, discussed in § 7.3.1, was utilised and trained on the 64 MICCAI images. The $256 \times 256 \times 3$ patch shown in Figure 14 was chosen for analysis as it contained a high density of nuclei from the NPC WSI – a patch with too few nuclei would not be informative. The patch was then fed into the U-Net to obtain a prediction in the form of a binary mask, of which N objects in the patch were nuclei (see Figure 14). Connected components labeling, a graph theory algorithm, was then applied to the mask to label all connected pixels that had a value of 255 with the numbers 1 through N. The 9 features listed in § 8.1 were extracted to create a feature vector of length 9 for each of these N nuclei, giving a 9-dimensional feature space.

8.3 Results of DBScan and k-means

DBScan clustering [54], discussed in § 7.2.1, was applied to the set of feature vectors. With ε varied and $minPts = 3$, the DBScan algorithm is visualised in Figure 13(a) which shows that the number of clusters reaches a maximum of four before it drops with increasing ε. One issue found with DBScan was that it categorised a large proportion of nuclei as noise with $minPts = 3$. It was not until $\varepsilon > 73$ when DBScan classifies less than 50% of the nuclei as noise, signaling that many of the predicted nuclei have irregular shape.



(a) DBScan (b) k-means

Figure 13: Cluster Analysis. (a) DBScan results. (b) k-means results.

k-means clustering [55], discussed in § 7.2.2, was then performed on the feature vectors to group the nuclei based on the 9 extracted features and into the optimum number of class, k. Figure 13(b) shows the graph of the number of clusters versus the corresponding total within cluster variance for our feature vectors. The 'elbow' of the graph occurs at around $k = 4$, agreeing with the result from DBScan. Hence, we utilise the

k-means algorithm to cluster our N nuclei into four clusters (see Figure 14).

The most determinative features were found through plotting heatmaps of the clusters versus the mean, variance and correlation of four of the features: size, eccentricity, intensity and standard deviation of intensity. Of the 9 features considered, the size and eccentricity were the most determinative how the nuclei were clustered (see Figure 15).

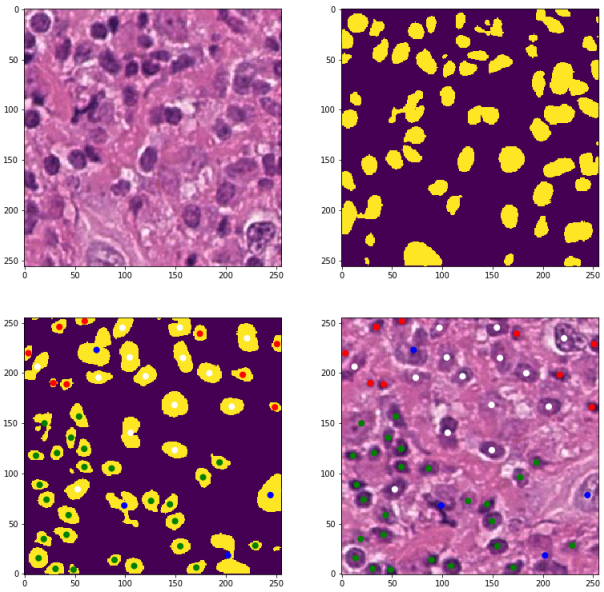


Figure 14: Clustering Prediction with k-means, k=4. Top left: patch from WSI, top right: U-Net prediction of nuclei, bottom left and right: k-means cluster prediction overlaid on the U-Net prediction and the original tile, respectively.

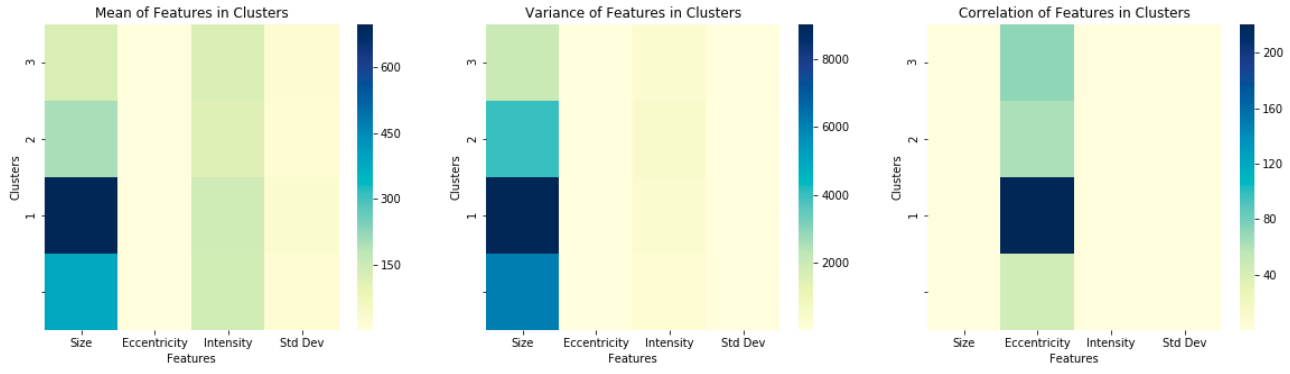


Figure 15: Heatmaps of 4 clusters versus 4 features for (left) mean, (centre) standard deviation and (right) correlation of the features.

9 Supervised Learning: Tissue Classification

After receiving two annotated NPC WSIs halfway into our project, we decided to investigate the level of accuracy attainable from utilising supervised learning methods, specifically, application of SVM and random forest classifiers, to perform tissue classification. We chose to apply SVM and random forest classifiers as our dataset containing the ground truth was small. When only a small dataset is available, a classifier trained on handcrafted features generally will outperform a CNN [56].

9.1 Data Preprocessing

Through generating masks from the two annotated WSIs, we extracted $256 \times 256 \times 3$ patches from each WSI into their respective folders, with patches in each further separated into folders based on tissue type. Using the first annotated image, shown in Figure 1, as the reference image, we applied a stain normalisation method proposed by Macenko *et al.* to the patches from the second annotated WSI [57].

Since the number of extracted patches for each WSI consisted of approximately $10,000$ patches, only a subset of patches of one WSI was utilised in the training phase for each model. The testing was then performed on both patches from the WSI used to train and patches from the other WSI.

9.2 Feature Extraction

Four types of features, the central moments, GLCM, LBP and perceptual features, were extracted for each patch, generating a 74-length feature vector. These features were utilised as they can describe the variation found in cancer tissue and give a characterization of the entire patch instead of individual objects within [56]. Thus, they are suitable for tissue classification.

9.3 Model Training and Testing

To measure the difference between a small and large dataset, we performed our training and testing in two runs: the first with a smaller training dataset and a second with a larger training dataset. In each run, SVM and random forest classifiers were implemented.

Moreover, since are feature vector was large compared to our available dataset, the method of principal component analysis (PCA) was utilised as well. Application of PCA to data may be beneficial when the dimension of data space is relative to the amount of data available [55]. As PCA reduces the input dimension significantly, it helps in avoiding overfitting [55]. After applying PCA, each feature vector was reduced to 3 components.

For the first run, the data sets utilised composed of 409 feature vectors each of both tumour and non-tumour patches of the first WSI (818 total) and 259 feature vectors each of both tumour and non-tumor patches of the second WSI (518 total). We then did a 350-59 split into training and testing sets, respectively, for both the tumor and non-tumour patches of the first WSI data. Each model was then trained on only those training sets from the first WSI.

As part of the first run, we also trained each model using all 818 feature vectors of the first WSI and reserved the 518 feature vectors of the second WSI for testing. Additionally, stain normalization was not applied to the patches of the second WSI for this first run.

In the second run, 1425 tumour and 940 non-tumour feature vectors were derived from the first WSI and 1099 non-tumour and 919 tumour feature vectors were derived from the second WSI. Macenko stain normalization was applied to the patches of the second WSI. Once again, for the first WSI, we split the data into a training set of 1225 tumour and 740 non-tumour feature vectors for training and 200 tumour and 200 non-tumour patches for testing.

9.4 Results

(A) SVM, Sample 1, 74 Features
(409/ 409) tumor/ non-tumor patches of WSI1
(259/ 259) tumor/ non-tumor patches of WSI2

Testing 1, (350/350) vs (59/59)	Precision	Recall	F1	Support
non-tumour	0.85	0.80	0.82	59
tumor	0.81	0.86	0.84	59
Average	**0.83**	**0.83**	**0.83**	**118**
Testing 1, PCA 3	Precision	Recall	F1	Support
non-tumour	0.54	0.42	0.47	50
tumor	0.52	0.64	0.58	50
Average	**0.53**	**0.53**	**0.52**	**100**
Testing 2, not-normed	Precision	Recall	F1	Support
non-tumour	0.43	0.20	0.27	259
tumor	0.48	0.74	0.58	259
Average	**0.46**	**0.47**	**0.43**	**518**
Testing 2, not-normed, PCA 3	Precision	Recall	F1	Support
non-tumour	0.62	0.58	0.59	259
tumor	0.60	0.64	0.62	259
Average	**0.61**	**0.61**	**0.61**	**518**
Testing 2, normed	Precision	Recall	F1	Support
non-tumour	0.53	0.19	0.27	259
tumor	0.51	0.83	0.63	259
Average	**0.52**	**0.51**	**0.45**	**518**
Testing 2, normed, PCA	Precision	Recall	F1	Support
non-tumour	0.62	0.46	0.53	259
tumor	0.57	0.72	0.64	259
Average	**0.60**	**0.59**	**0.58**	**518**

(A') Random Forest, Sample 1, 74 Features
(409/ 409) tumor/ non-tumor patches of WSI1
(259/ 259) tumor/ non-tumor patches of WSI2

Testing 1, (350/350) vs (59/59)	Precision	Recall	F1	Support
non-tumour	0.93	0.85	0.88	59
tumor	0.86	0.93	0.89	59
Average	**0.89**	**0.89**	**0.89**	**118**
Testing 1, PCA 3	Precision	Recall	F1	Support
non-tumour	0.47	0.42	0.44	50
tumor	0.47	0.52	0.50	50
Average	**0.47**	**0.47**	**0.47**	**100**
Testing 2, not-normed	Precision	Recall	F1	Support
non-tumour	0.66	0.89	0.76	259
tumor	0.84	0.55	0.66	259
Average	**0.75**	**0.72**	**0.71**	**518**
Testing 2, not-normed, PCA 3	Precision	Recall	F1	Support
non-tumour	0.60	0.58	0.59	259
tumor	0.59	0.61	0.60	259
Average	**0.59**	**0.59**	**0.59**	**518**
Testing 2, normed	Precision	Recall	F1	Support
non-tumour	0.71	0.78	0.74	259
tumor	0.76	0.68	0.72	259
Average	**0.73**	**0.73**	**0.73**	**518**
Testing 2, normed, PCA	Precision	Recall	F1	Support
non-tumour	0.64	0.53	0.58	259
tumor	0.60	0.71	0.65	259
Average	**0.62**	**0.62**	**0.61**	**518**

(B) SVM, Sample 2, 74 Features
(1425/ 940) tumor/ non-tumor patches of WSI1
(919/ 1099) tumor/ non-tumor patches of WSI2

Testing 1, (1225/740) vs (200/200)	Precision	Recall	F1	Support
non-tumour	0.94	0.81	0.87	200
tumor	0.83	0.95	0.89	200
Average	**0.89**	**0.88**	**0.88**	**400**
Testing 1, PCA 3	Precision	Recall	F1	Support
non-tumour	0.59	0.43	0.50	200
tumor	0.55	0.69	0.62	200
Average	**0.57**	**0.56**	**0.56**	**400**
Testing 2, normed	Precision	Recall	F1	Support
non-tumour	0.61	0.45	0.52	1,099
tumor	0.50	0.66	0.57	919
Average	**0.56**	**0.55**	**0.54**	**2,018**
Testing 2, normed, PCA	Precision	Recall	F1	Support
non-tumour	0.88	0.49	0.63	1,099
tumor	0.60	0.92	0.73	919
Average	**0.75**	**0.69**	**0.68**	**2,018**

(B') Random Forest, Sample 2, 74 Features
(1425/ 940) tumor/ non-tumor patches of WSI1
(919/ 1099) tumor/ non-tumor patches of WSI2

Testing 1, (1225/740) vs (200/200)	Precision	Recall	F1	Support
non-tumour	0.95	0.82	0.88	200
tumor	0.85	0.95	0.90	200
Average	**0.90**	**0.89**	**0.89**	**400**
Testing 1, PCA 3	Precision	Recall	F1	Support
non-tumour	0.32	0.18	0.23	200
tumor	0.43	0.61	0.50	200
Average	**0.37**	**0.40**	**0.37**	**400**
Testing 2, normed	Precision	Recall	F1	Support
non-tumour	0.62	0.64	0.63	1,099
tumor	0.55	0.53	0.54	919
Average	**0.59**	**0.59**	**0.59**	**2,018**
Testing 2, normed, PCA	Precision	Recall	F1	Support
non-tumour	0.87	0.46	0.60	1,099
tumor	0.59	0.92	0.72	919
Average	**0.74**	**0.67**	**0.65**	**2,018**

Figure 16: SVM and Random Forest Tissue Classification Results. Top row: (A), (A') Results based on a small dataset. Bottom row: (B), (B') Results based on a large dataset

The results from our models are shown in Figure 16. The top two tables show the results when we used the smaller dataset whilst the bottom two tables show the results when we increased the dataset. The first two rows of each table show precision, recall and F1 values for the classes, tumour and non-tumour, whilst the last row shows the average of both. Lastly, the support column indicates the sample size.

The results for the SVM model F1-score were 0.82 for non-tumour, 0.84 for tumour and 0.83 combined. The random forest model gave slightly better F1-scores with 0.88 for non-tumour, 0.89 for tumour and 0.89 combined. Application of PCA with 3 principal components actually reduced the F1-score for both the SVM and the random forest model for testing 1. With PCA, the F1-score for the SVM model decreased to 0.47 for non-tumour, 0.58 for tumour and 0.52 combined. The performance of the random forest model was better at 0.44 for non-tumour, 0.50 for tumour and 0.47 combined.

Although the generalization performance improved for the SVM model with PCA in testing 2 from 0.27 for non-tumour, 0.58 for tumour and 0.43 combined, respectively to 0.59 for non-tumour, 0.62 for tumour and 0.61 combined, respectively, it decreased the performance for the random forest from 0.76 for non-tumour, 0.66 for tumour and 0.71 combined, respectively to 0.59 for non-tumour, 0.60 for tumour and 0.59 combined. The F1-score improved only slightly by using the Macenko stain normalization method for both the SVM model and the random forest model.

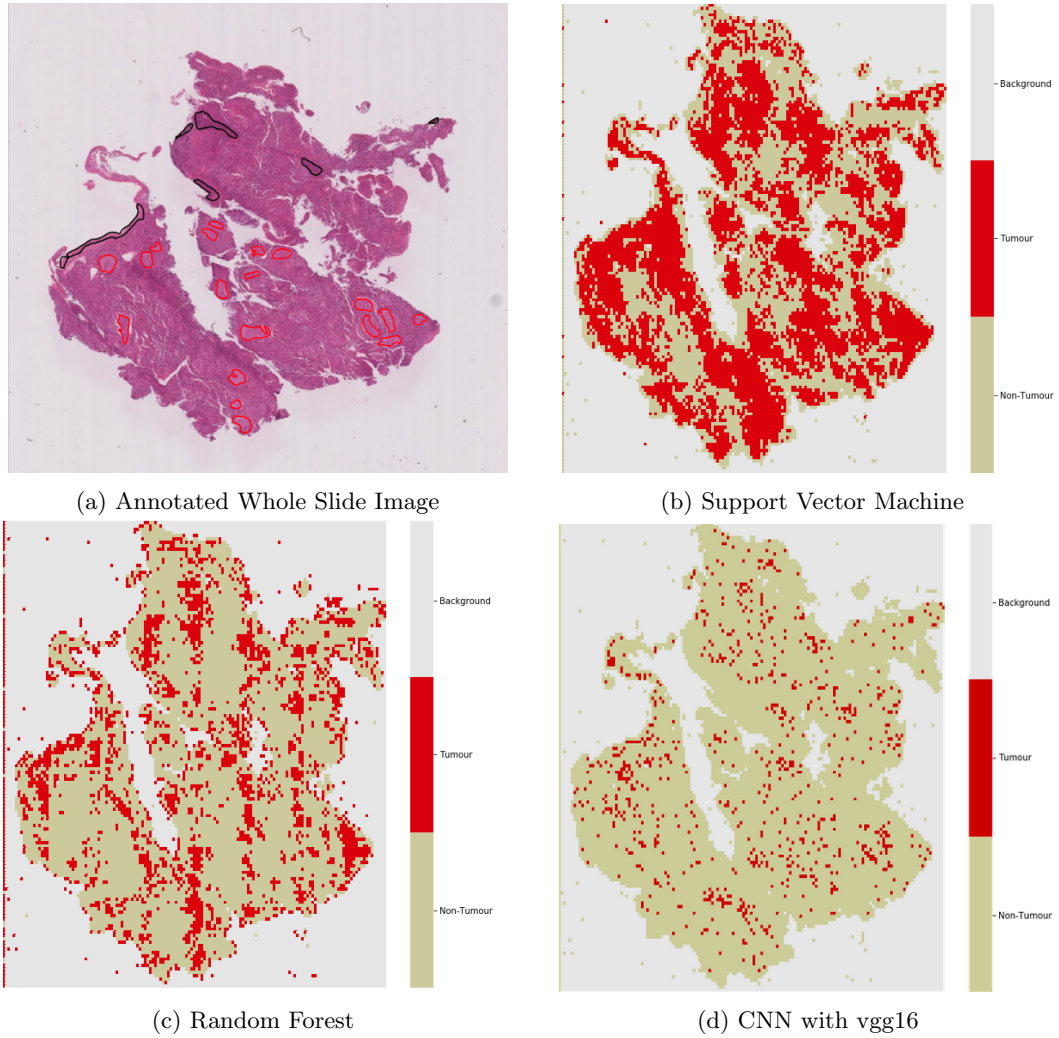
Increasing the size of the data improved SVM F1-score slightly to 0.87 for non-tumour, 0.89 for tumour and

0.88 combined. Random forest F1-score was relatively the same at 0.88 for non-tumour, 0.90 for tumour and 0.89 combined. Again, applying PCA reduced the performance in both cases.

For the second test of this run, SVM F1-score improved slightly to 0.52 for non-tumour, 0.57 for tumour and 0.54 combined whilst random forest F1-score fell to 0.63 for non-tumour, 0.54 for tumour and 0.59 combined.

Application of PCA improved the generalization performance for both models as SVM F1-score increased to 0.63 for non-tumour, 0.73 for tumour and 0.68 combined whilst random forest F1-score increased to 0.60 for non-tumour, 0.72 for tumour and 0.65 combined.

9.5 Predicted Heatmaps



(a) Annotated Whole Slide Image

(b) Support Vector Machine

(c) Random Forest

(d) CNN with vgg16

Figure 17: Tumour and non-tumour regions in (a) NPC tissue annotated by pathologist and prediction with (b) support vector machine, (c) random forest (d) vgg-16 pre-trained on image net with final three layers trained on NPC tumour and non-tumour patches.

In order to be medically informative, pathologists must be able to visually see the results of WSI tissue classification. Accordingly, using the features discussed in §9.2, we obtained feature vectors from patches generated by the first annotated WSI to train SVM and random forest classifiers. We also used a vgg-16 CNN classifier pre-trained on ImageNet with the final 3 layers trained on patches from the first WSI. The vgg-16 model was discussed in §3.4 & §7.3.1. Testing was done on patches comprising the entire second

annotated WSI, providing for a visual comparison of the accuracy. Figure 17 shows the results of our heatmap predictions for tumour and non-tumour regions in combination with Otsu thresholding. As is visually apparent the predictions are of low accuracy.

10 Discussion

10.1 Unsupervised Learning

In performing unsupervised learning of the nucleic features, the most noticeable issue perhaps was that nuclei close in proximity were predicted as one nuclei by the U-net, leading to inaccurate clustering at the end. With better nuclei segmentation, the accuracy of the clustering algorithms would vastly improve. This could be done by training the U-Net with more images and masks or by segmentation techniques such as watershed separation [53].

Another issue was scale. As a patch from only a specific area of the NPC WSI was used, a better result might be obtained by including a large set of patches from different regions of the tissue into the analysis. This could be done through extracting feature vectors of predicted nuclei from different tissue regions and then concatenating them at the end to perform clustering analyses. However, this would become computationally heavy if thousands of nuclei are predicted. Moreover, overlaying the resulting labels for thousands of nuclei would be difficult and would not provide a clear visual of the clustering.

The nine features utilised were object-level or morphological features based on the size and shape of the nuclei and are not exhaustive. The features extracted could be extended to include other object-level features, such as the length of the perimeter or compactness of the nuclei, and spatial-arrangement features, based on graph theory, that consider the location of nuclei relative to its neighbors [21, 53]. These extensions would be of interest to further work.

10.2 Supervised Learning

The results displayed in Figure 16 show that the random forest model performed slightly better than the SVM. Examining Figure 17 to compare the SVM and random forest results to a CNN, the SVM over predicts for tumour whilst the CNN under predicts for tumour. Though the random forest classifier appears to achieve the best performance out of the three models, the disadvantage in its application is that it has a tendency to overfit and provides no knowledge regarding which features were most determinative [55]. On the other hand, the benefit of utilising an SVM is that it does allow for this information to be attained [55].

Significantly, for both the SVM and random forest classifiers, our testing showed that training with data from one WSI and testing on data from another produced a low accuracy. On the contrary, when data from the same WSI was used in both training and testing, the accuracy was high. Although there are other textural and morphological features that we could add to try and improve our models, as the textural features we chose have previously produced accurate tissue classification [56], this would seem to indicate that in our case, the features and patterns in one WSI are much different from the other. NPC tissue is characterized by a high degree of heterogeneity which may account for this. Moreover, features extracted from one patient's tissue may not be completely representative of another patient's.

Additionally, as we only had two annotated WSIs, poor prediction accuracy should not be surprising. The papers we considered during this project all had larger datasets [18, 56]. As each patient may have characteristics that are distinct, a larger dataset representing a population of NPC patients would better capture variations in tissue samples and likely produce a more accurate result. Thus, the amount of annotated WSI data that we had was a considerable limitation on our computational analyses. In view of this, the amount of digitised data available will be critical for digital pathology to be a success.

11 Conclusion

In this project, we explored how pathology labs are employing machine learning techniques, specifically through digital pathology, to enhance the diagnoses of disease and improve therapies. Through conducting a feasibility analysis, we determined that though some pathology labs and governments may be reluctant to implement digital pathology, they are optimistic that it will be the future of pathology once its success has been proven. Moreover, investigating digital pathology in the context of nasopharyngeal carcinoma, we found that there is great potential for digital pathology to not only improve diagnoses but reveal the mechanisms that cancer cells assume to overwhelm the immune system – which may lead to a complete cure. After implementing two different techniques, an unsupervised and a supervised method, we learned that having sufficient data was necessary for accurate prediction and classification of tissue samples. Finally, through this project, we aimed to show that there is much more potential to machine learning than classifying images of cats and dogs – perhaps it is the key to solving cancer.

12 Acknowledgements

We would like to thank Professor Nasir Rajpoot for his time, resources and suggestions for this project. We would also like to express our gratitude to his lab, with particular mention to Jev, Peter, Angus, Simon, Kat and Shaban for their discussions and assistance.

References

[1] Pantanowiz, L. et al. *Digital Pathology.* American Society for Clinical Pathoogy, 2017.

[2] Cancer Research UK, *Nasopharyngeal cancer.* http://www.cancerresearchuk.org/about-cancer/nasopharyngeal-cancer, Accessed on 11/06/2018.

[3] Helliwell, T. *Diagnostic digital pathology strategy.* The Royal College of Pathologists, 2017.

[4] Williams, B. et al. *Future-proofing pathology: the case for clinical adoption of digital pathology.* Journal of Clinical Pathology, 2017.

[5] Cross, S. et al. *Best practice recommendations for implementing digital pathology.* The Royal College of Pathologists, 2018.

[6] Bankhead, P. et al. *QuPath: Open source software for digital pathology image analysis.* Nature scientific reports, 2017.

[7] Carpenter, A. et al. *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome biol. 2006.

[8] Sirinukunwattana, K. et al. *Gland segmentation in colon histology images: The glas challenge contest.* arXiv.org ePrint archive, 2016.

[9] Bejnordu B. et al. *Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer.* JAMA, 2017

[10] Trahearn, N. et al. *Multi-class stain separation using independent component analysis.* Medical imaging 2015: Digital PAthology, SPIE Proceedings Vol. 9420, 2015

[11] Wang, M. et al. *A patch-based tensor decomposition algorithm for M-FISH image classification.* Cytometry A, 2017.

[12] Roy, P. et al. *Adaptive thresholding: A comparative study.* ICCICCT 2014 international conference proceedings, IEEE, 2014.

[13] Romo-Bucheli, D. et al. *A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers.* Cytometry A, 2017.

[14] Wu, H. et al. *Segmentation of intestinal gland images with iterative region growing.* Journal of Microscopy, Vol. 220, 2005.

[15] Sirinukunwatta, K. et al. *A stochastic polygons model for glandular structures in colon histology images.* IEEE Transactions on medical imaging, 2015.

[16] Paulik, R. et al. *An Optimized Image Analysis algorithm for detecting nuclear signals in digital whole slides for histopathology.* Cytometry A, 2017.

[17] The ImageNet image database: http://www.image-net.org, last accessed 11.6.18.

[18] Wang, D. et al. *Deep learning for identifying metastatic breast cancer.* arXiv.org ePrint archive, 2016.

[19] Litjens, G. et al. *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Nature scientific reports, 2016.

[20] Zhang, Y. et al. *Deep adversarial networks for biomedical image segmentation utilizing unannotated images.* MICCAI 2017 conference proceedings, Springer Verlag, 2017

[21] Kumar, N. et al. *A dataset and a technique for generalized nuclear segmentation for computational pathology.* IEEE Transactions on medical imaging, 2017.

[22] Liu, Y. et al. *Detecting cancer metastases on gigapixel pathology images.* arXiv.org ePrint archive, 2017.

[23] Paeng, K. et al. *A unified framework for tumor proliferation score prediction in breast histopathology.* arXiv.org ePrint archive, 2017.

[24] Qaiser, T. et al. *HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues.* Histopathology, 2017.

[25] Alsubale, N. et al. *Stain deconvolution using statistical analysis of multi-resolution stain colour representation.* PLoS One, 2017.

[26] Lindberg, A. et al. *Quantitative tumor heterogeneity assessment on a nuclear population basis.* Cytometry A, 2017.

[27] Valkonen, M. et al. *Metastasis detection from whole slide images using local features and random forests.* Cytometry A, 2017.

[28] Walker, A. et al. *Perceptions of pathology informatics by non-informaticist pathologists and trainees.* Journal of pathology informatics, 2016

[29] Williams, B. et al. *Digital pathology access and usage in the UK: results from a national survey on behalf of the National Cancer Research Institute's CM-Path initiative.* Journal of clinical pathology, 2018.

[30] Bissell, M. J., Hines, W. C., *Why dont we get more Cancer? A Proposed Role of the Microenvironment in Restraining Cancer Progression.* Nature Medicine, 2008.

[31] Mbeunkui, F., Johann Jr., D. J., *Cancer and the Tumor Microenvironment: A Review of an Essential Relationship.* Cancer Chemother Pharmacol., 2009.

[32] Kallert et al., *Replicating Viral Vector Platform Exploits Alarmin Signals for Potent CD8+ T-cell-mediated Tumour Immunotherapy.* Nature Communications, 2017.

[33] Whiteside, T. L., *The Tumor Microenvironment and its Role in Promoting Tumor Growth.* Oncogene, 2008.

[34] Balkwill, F. R., Capasso, M., Hagemann, T., *The Tumor Microenvironment at a Glance*. Journal of Cell Science, 2012.

[35] Hirata, E., Sahai, E., *Tumor Microenvironment and Differential Responses to Therapy*. Cold Spring Harbor Perspectives in Medicine, 2017.

[36] Maman, S., Witz, I.P., *A History of Exploring Cancer in Context*. Nature Reviews, 2018.

[37] Huang et al., *Interplay of Viral Infection, Host Cell Factors and Tumor Microenvironment in the Pathogenesis of Nasopharyngeal Carcinoma*. Cancers, 2018.

[38] Gourzones et al., *Cellular Interactions in Nasopharyngeal Carcinomas*. Nasopharyngeal Carcinoma: Keys for Translational Medicine and Biology, 2013.

[39] Wang et al., *BART miRNAs: An Unimaginable Force in the Development of Nasopharyngeal Carcinoma*. European Journal of Cancer Prevention, 2017.

[40] Heindl, A, Nawaz, S, Yuan, Y, *Mapping Spatial Heterogeneity in the Tumor Microenvironment: A new era for Digital Pathology*. Pathobiology in Focus, 2015.

[41] Balsat et al., *Improved Computer-assisted Analysis of the Global Lymphatic Network in Human Cervical Tissues*. Modern Pathology, 2014.

[42] Feichtenbeiner et al., *Critical Role of Spatial Interaction Between CD8+ and Foxp3+ Cells in Human Gastric Cancer: The Distance Matters*. Cancer Immunol Immunother, 2014.

[43] Nawaz, S., Yuan, Y., *Computational Pathology: Exploring the Spatial Dimension of Tumor Ecology*. Cancer Letters, 2016.

[44] Wells et al., *Spatial and Functional Heterogeneities Shape Collective Behavior of Tumor-Immune Networks*. PLOS Computational Biology,2015.

[45] Peruzzi, P., Chiocca, E. A., *Understanding Glioma Invasion: A Necessity for Effective Therapy*, World Neurosurgery, 2013.

[46] Koos et al., *Next-Generation Pathology Surveillance of Tumor Microecology*. Journal of Microbiology, 2015

[47] Koury et al., *Immunotherapies: Exploiting the Immune System for Cancer Treatment*. Journal of Immunology Research, 2018.

[48] Zacharakis et al., *Immune Recognition of Somatic Mutations Leading to Complete Durable Regression in Metastatic Breast Cancer*. Nature Medicine, 2018.

[49] Farahani, N., Parwani, A., Liron P., *Whole slide imaging in pathology: advantages, limitations, and emerging perspectives*. Dove Press Journal, Pathology and Laboratory Medicine International, 2015.

[50] Ronneberger O., Fischer P. and Brox T.. *U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597v1 [cs.CV]*, 2015.

[51] Simonyan K., Zisserman A., *Very Deep Convolutional Networks for Large Scale Image Recognition. arXiv:1409.1556*, 2015.

[52] Thompson L. D. R. *Update on Nasopharyngeal Carcinoma*. Head and Neck Pathology. 2007;1(1):81-86. doi:10.1007/s12105-007-0012-7.

[53] Irshad H., Student Member, IEEE, Veillard A., Roux L., and Racoceanu D., Member, IEEE *Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review - Current Status and Future Potential*. IEEE Reviews in Biomedical Engineering, VOL. 7, 2014.

[54] Ienco D., Bordogna G., *Fuzzy Extensions of the DBScan Clustering Algorithm*. Soft Computing, Springer Verlag, 2016.

[55] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Data Mining, Inference and Prediction. Ch. 14.3, Springer, Second Edition, 2008.

[56] Kather J. N. et al., *Multi-class texture analysis in colorectal cancer histology.* Nature, Scientific Reports, 2016.

[57] Macenko M. et al., *A method for normalizing histology slides for quantitative analysis.* IEEE, 2009

[58] Mitchinson, M. et al. *Essentials of Pathology.* Wiley-Blackwell, 1995.

Appendix

A Foundational pathology

Cells have a life cycle. They are born through the division of a parent cell during mitosis and they die through the calculated destruction of apoptosis. This cycle allows living organisms to grow and survive cell defects by controlling these event rates. Diseases however may perturb this delicate balance causing an abnormal swelling of the tissue, a neoplasm, affecting the surrounding healthy tissues. Cancer is a family of diseases that manifest from a corruption of the genetic code within a cell commonly caused by exposure of mutation causing carcinogens or ionising radiation. The many forms of cancer fall within this set of growth producing diseases [58].

A.1 Neoplasms

There are two forms of neoplasm which affect the grading of the cancer; benign and malignant. Benign neoplasm may form a tumour composed of cells that are well differentiated, acting in a similar way to the healthy, undiseased cell but more numerous. The mass forms a geometry related to the flexibility of the surrounding substrate tissue, tumours may be round or extend into cavities or vessels. The tissue of a benign neoplasm may be differentiated as the corresponding healthy tissue with ordered cells and expected cell correlations. Additionally, if the cells of the tissue function by secreting biosynthesised molecules then these secretions may continue. Therefore most medical complications caused by benign tumours are a consequences of the high cell density and physical tumour size. The physical size of the mass builds a physical pressure that tends to degrade the surrounding tissue. This becomes more dangerous if this pressure builds upon a blood vessel causing a restriction of blood flow. If this grows within tissue membranes the disrupted growth may perforate the membrane leading to ulceration and further complications. Functional cells producing secretions may upset the homeostatic control of the organism, such as a tumour on the adrenal gland uncontrollably secreting adrenaline. Benign tumours may also become malignant over time.

Malignant tumours are usually not so well differentiated. They tend to be less well organised with irregular ordering. The resulting diseased tissue may then appear anomalous under a microscope. Malignant cells infiltrate the surrounding tissue clusters commandeering their blood supplies. If the tumour metastasises, masses may grow a relatively large distance from the original tumor. These new masses form neoplasms which may metastasis themselves. One method of transportation is to infiltrate a lymph node and use the bodies own circulation systems to settle in new parts of the organism. The malignant tumour may also cause complications akin to the benign tumour but tend to be more aggressive in their behaviour. The chemical secretions from malignant cells may themselves be anomalous which may be detected in the tissue sample providing one means of distinguishing benign from malignant [58].

A.2 Nasopharyngeal carcinoma

One form of cancer is Nasopharyngeal carcinoma (NPC). NPC is an example of a head and neck cancer forming in the nasopharynx, a cavity behind the mouth and at the bottom of the nasal cavity. NPC doesnt affect all populations equally. More cases are found in males rather than females and most cases come from ethnic Chinese populations. This has the impact of NPC being relatively rare in the UK with only approximately 250 cases a year. The common trigger for the development of NPC is through infection of

the Epstein-Barr virus (EBV). This link was determined through the identification of EBV DNA within the genome of cells in NPC tumours, however a virus is unable to cause the formation of a tumour and the true trigger remains elusive. The identification and classification of a carcinoma may be performed through a histopathological analysis [58, 2].

B Histopathology

Cancerous tissues may be studied as solid tissue sections, termed histopathology. These tissue samples are obtained through biopsies producing either small samples of suspected tumours or whole diseased organs. These samples go through a series of processing stages to diagnose the disease. These stages include preprocessing, slide preparation, and slide evaluation. These stages constitute an intensely manual process rarely seen in other pathological methodologies [1].

B.1 Sample preprocessing

After a suspected diseased tissue region is identified a sample of the material is taken during a biopsy. This material is then sent to a pathology lab alongside a pathology request form. The pathology request form contains information such as the patients details, the clinicians diagnosis regarding the sample, as well as the symptoms expressed by the patient. The sample and request form as received by the lab and catalogued into a case file. The tissue sample is then processed.

The specimens are stored in a formalin solution. Formalin is an aqueous solution of formaldehyde which is a toxic carcinogen itself. Therefore the specimens are sealed and stored in labelled containers awaiting processing. When needed, the specimen are loaded into cassettes, cutting if necessary. For larger samples multiple cassettes may be needed as well as the removal of non-diseased tissue segments. Biological tissues have a high water content. Water would provide complications downstream in the processing. Therefore the water is to be removed.

The cassettes are immersed in a chemical dehydration bath. The aim of this process is to replace the water content of the tissue with paraffin wax through a series of chemical substitutions. Firstly, the cassettes undergo bathing in varying concentrations of aqueous ethanol to drive off the water. Then they are treated with a hydrophobic clearing agent, xylene, which fixes the tissue. Liquid Paraffin wax is then added which solidifies the samples upon cooling at room temperature. Different tissues have different water compositions, for instance fatty tissues tend to have a greater water content in the adipose tissue, requiring longer processing times. This process is time intensive varying from 4 hours for a rapid turn around ccle to up to 13 hours for a standard process, typically an overnight venture.

B.2 Slide preparation

The wax impregnated cassettes are used to produce the pathological slides. The cassettes are loaded onto paraffin blocks to aid the manual cutting. From this block thin veneers of the slide, of the order of a few microns thick, are produced. These veneers are floated on a heated water bath to soften the wax and enable the tissue to flatten on the smooth water surface. This is done to minimise the folding in the tissue samples. These samples are then placed onto a glass slide and allowed to cool. These slides are then dried to ensure a firm attachment of the tissue to the slide before labelling and attaching the slide to the patient case file. Next the slides undergo pathological staining. The choice of staining required would be detailed on the pathology request form with the standard staining being Hematoxylin and Eosin (H&E) staining. Hematoxylin binds to negatively charged biomolecules such as the DNA in the cell nucleus, producing a dark blue colour. Meanwhile, Eosin binds to positively charged biomolecules such as the cells cytoplasm, producing a red colour. This staining combination reveals the cell volumes (red) and nuclei (blue), although not the cell membrane, to the naked eye. H&E staining is performed automatically using a staining machine. The quantity of stain used may be subject to the pathologists preference producing inter-pathologist slide stain intensity variations. The stains themselves may also vary depending on the manufacturer leading to interdepartmental variation. The slides then go through quality assurance and are enclosed within the patient case file alongside the remnants of the specimen cassette block. This specimen block is kept in case the pathologist requests further staining and new slides are required to be made.

B.3 Evaluation

Slides are evaluated by pathologists; either physically under a microscope or digitally using a digitised slide on a high resolution screen. The pathologists will use the patient case file in conjunction with the slides to grade the tissue sections for diseases. For example, grading criteria may involve the number of mitotic cells under certain magnifications, visible as deep blue spots. The grade may then be used to suggest courses of treatment for the patient. As the treatment is dependent on the pathologists report any uncertainty in the grade may be eased through the use of an additional stain mode. These additional stains may be for immunohistochemical staining (IHC) which looks for specific proteins in the cells or specialist biomarkers to probe bespoke features. These additions would require the specimen block to be reprocessed to produce new slides to inspect.