Computational Analysis of Tumour Microenvironment in mIHC Stained Diffuse Glioma Samples

Master's Thesis
University of Turku
Data Analytics
2022
B.Sc. Jacob Costa
Examiners:
  Prof. Jyrki Piilo
  Prof. Pekka Ruusuvuori

UNIVERSITY OF TURKU
Department of Physics and Astronomy

Costa Jacob Computational Analysis of Tumour Microenvironment in mIHC Stained Diffuse Glioma Samples

Master's Thesis, 61 pp.
Data Analytics
May 2022

Healthcare is a sector that has been notoriously stagnant in digital innovation, nevertheless its transformation is imminent. Digital pathology is a field that is being accentuated in light of recent technological development. With capacity to conduct high-resolution tissue imaging and managing output digitally, advanced image analysis and Machine Learning can be subsequently applied. These methods provide means to for instance automating segmentation of region-of-interests, diagnosis and knowledge discovery.

Brain malignancies are particularly dire with a high fatality rate and relatively high occurrence in children. Diffuse gliomas are a subtype of brain tumours whose biological behavior range from very indolent to extremely aggressive, which is reflected in grading I - IV. The brain tumour micro-environment (TME) — local area surrounding cancerous cells with a plethora of immune cells and other structures in interaction — has emerged as a critical regulator of brain tumour progression. Researchers are interested in immunotherapeutic treatment of brain cancer, since modern approaches are insufficient in treatment of especially the most aggressive tumours. Additionally, the TME is rendered difficult to understand. Multiplex Immunohistochemistry (mIHC) is a novel approach in effectively mapping spatial distribution of cell types in tissue samples using multiple antibodies.

In this thesis, we investigate the TME in diffuse glioma mIHC samples for three patient cases with 2-3 differing tumour grades per patient. From the 18 possibilities we selected 6 antigens (markers) of interest for further analysis. In particular, we are interested in how relative proportion of positive antigens and mean distance to nearest blood vessel vary for our selected markers in tumour progression. In order to acquire desired properties, we register each corresponding image, detect nuclei, segment cells and extract structured data from region channel intensities along with their location and distance to nearest blood vessel.

Our primary finding is that M2-macrophage and T cell occurrence proportions as well as their mean distance to blood vessel grow with increasing tumour grade. The results could suggest that aforementioned cell types are of low quantity in near vicinity of blood vessels in low tumour grades, and conversely with higher quantities and more homogeneous distribution in aggressive tumours. Despite the several potential error sources and non-standardized processes in the pipeline between tissue extraction and image analysis, our results support pre-existing knowledge in that M2-macrophage proportion has a positive correlation with tumour grade.

Keywords: Digital Pathology, Computational Pathology, Bioimage Analysis, Advanced Image Analysis, Multiplex Immunohistochemistry, Tumour Microenvironment

TURUN YLIOPISTO
Fysiikan ja Tähtitieteen laitos

Costa Jacob Kasvainmikroympäristön laskennallinen analyysi diffuusigliooma mIHC-kudosnäytteissä

Pro gradu -tutkielma, 61 pp.
Data analytiikka
Toukokuu 2022

Terveydenhuollon digitaalinen kehitys on ollut hidasliikkeistä muihin sektoreihin verrattuna. Tästä huolimatta, terveydenhuollon digitaalinen muunnos on välitön ja asiaan liittyvä tutkimus jatkuvaa. Digitaalinen patologia on ala, joka viime aikaisen teknologisen kehityksen myötä on korostunut. Kudoskuvantaminen korkealla resoluutiolla ja näytteiden digitaalinen hallinta on mahdollistanut kehittyneen kuvanalysiin sekä koneoppimisen soveltamisen. Nämä metodit luovat keinot esimerkiksi biologisesti merkittävien alueiden segmentointiin, diagnoosiin ja uuden tieteellisen tiedon tuottamiseen.

Aivokasvaimet ovat järkyttäviä, sillä tapauskuolleisuus ja esiintymä nuorissa ovat suhteellisen korkealla. Diffuusigliomat ovat aivokasvainten alatyyppi, jonka sisältämät kasvaimet luokitellaan niiden aggressiivisuuden perusteella eri graduksiin väliltä I - IV. Kasvaimen mikroympäristö (TME), eli syöpäsolujen paikallinen ympäristö sisältäen mm. runsaasti immuunipuolustuksen soluja vuorovaikutuksessa, on osoittautunut merkittäväksi tekijäksi kasvaimen kehityksen suhteen. Aivosyövän tutkimus painottuu immunoterapeuttisiin ratkaisuihin, sillä nykyiset hoitomuodot eivät ole tarpeeksi tehokkaita etenkään kaikista aggressiivisimpien kasvainten hoidossa. Lisäksi mikroympäristö voi olla vaikea ymmärtää. Monikanavainen immunohistokemiallinen värjäys (mIHC) on uudenlainen lähestymistapa solutyyppien spatiaalijakauman kartoittamiseen kudosnäytteissä tehokkaasti hyödyntäen useita vasta-aineita.

Tässä opinnäytetyössä tutkitaan diffuusigliooma mIHC-näytteitä kolmelle potilastapaukselle. Jokaista potilasta kohti on 2-3 näytettä eri kasvainlaaduista ja yhteensä 18 mIHC-kanavaa per näyte, joista 6 otettiin tarkasteluun. Tarkalleen ottaen, solutyyppien aktivaatioiden osuudet positiivisten antigeenien perusteella ja keskimääräinen etäisyys lähimpään verisuoneen jokaista ryhmää kohti lasketaan eri kasvaimen laaduissa. Tavoitteen saavuttamiseksi näytteitä vastaavat kuvat rekisteröidään, tumat tunnistetaan, solualueet segmentoidaan ja kerätään jäsenneltyä tietoa alueiden intensiteettikanavista mukaan lukien sijainti ja sijaintia vastaava etäisyys lähimpään verisuoneen.

Pääasiallinen löytö on, että M2-makrofagien ja T-solujen suhteelliset osuudet sekä keskimääräinen etäisyys lähimpään verisuoneen nousevat kasvaimen ollessa aggressiivisempi. Tulokset saattavat ehdottaa, että edellämainitut solutyypit ovat vähäisiä ja verisuonten lähellä kun kasvain on hyvänlaatuinen ja vastaavasti suurimilla osuuksilla ja enemmän homogeenisesti jakautunut kun kasvain on aggressiivisempi. Useista virhelähteistä ja kudosanalyysin liittyvistä ei-standardisoiduista prosesseista huolimatta, tuloksemme tukevat ennaltatiedettyä tietoa siitä, että M2-makrofagien osuudella on positiivinen korrelaatio kasvaimen laatuun.

Avainsanat: Digitaalinen patologia, laskennallinen patologia, biokuvainformatiikka, monikanavainen immunohistokemiallinen värjäys, kasvainmikroympäristö

Acknowledgements

This thesis topic would not have been possible without the help of others. An introductory course in basic bioimage informatics sparked an immediate interest in the intriguing world of tissue investigations in histopathology. Course teacher PhD student Laura Mairinoja kindly answered my inquiries about the field and helped me find a thesis topic by providing contact to Assistant Professor Pekka Ruusuvuori. To my requests of brain related topics, Ruusuvuori acquired me a thesis setting accordingly. Thank you to both.

Greatest gratitude regarding thesis completion goes to supervisors Masi Valkonen and Aliisa Tiihonen. Valkonen guided in all technical matters — from teaching me how to control the clusters to advising on suitable methods and implementing some pre-requisites to save me time. I only started learning Linux and writing bash scripts in the start of the thesis and now I would not manage without it. The thesis was built completely on work done by Tiihonen. Her work in using novel methods of multiplex staining in glioma tissue allowed me to analyze tremendously interesting tissue landscapes with a feeling that it was something new and rather unexplored. Additional thank you to Professor Jyrki Piilo for acting as my official supervisor from the Department of Physics and making sure I graduate.

Thank you to Tampere University for providing me a resource agreement for the acquisition of data and access to compute cluster Narvi. Thank you to the University of Turku and particularly departments of Physics and Computing for providing quality education and an analytical toolbox suitable for the future. Final gratitude goes to my brother Jonathan for providing belief in oneself over the years and mother for keeping order in my life.

Contents

Introduction

The digital transformation in healthcare has emerged from the increasing demand for chronic disease management, technological developments and the need to improve patient outcomes with scalable and cost-effective solutions [1]. This transformation is actualized by the integration of a rich data foundation with technologies such as Internet of Things (IoT), advanced analytics and Machine Learning (ML). Whilst other economic sectors have successfully undergone dramatic digitization, healthcare is lagging, particularly in transformation of *delivery* [2].

Digital pathology is a constantly evolving field that deals with the acquisition, management, sharing and interpretation of pathology information in a digital environment. The virtualization of traditional glass slides at high resolution — Whole Slide Imaging (WSI) – is at the forefront of modern digital pathology and is utilized in clinical as well as non-clinical settings [3]. In addition to basic functions such as image archiving and sharing, WSI enables the application of advanced image analysis as well as Artificial Intelligence (AI), extending the pathologists view beyond human perception [4].

Brain and other central nervous system (CNS) tumours are constituted by a large collection of histologically distinct subtypes. Although these tumours are rare, they compose a high fatality rate and thus a disproportionate effect on cancer mortality; two-thirds of patients die within a 5 year period after diagnosis. In recent years there has been development in the molecular understanding, classification, detection and diagnosis of these tumours, yet the causal factors remain largely unknown. [5]

Over the past few decades, the brain tumour microenvironment (TME) has been become recognized as a key regulator of cancer progression in brain malignancies. These microenvironments are highly complex and interconnected – consisting of a plethora of cell types in interaction — from various types of immune cells to the usual organ-residents such as neurons and astrocytes. Insights on the tumour-promoting roles of TME components have led to discovery of multiple potential therapies, however there is substantial progress to be made and countless open questions to address. [6]

Multiplex immunohistochemistry (mIHC) has emerged as an efficient approach to mapping spatial distributions of cell types in tumour microenvironments. In general, immunohistochemistry uses antibodies to detect certain antigens (markers) in a sample of tissue. Multiplex IHC facilitates simultaneous detection of several markers per tissue sample and subsequent analysis of tissue composition and cell-cell interactions — giving foundation to reproducible, efficient and cost-effective tissue explorations. [7]

In this thesis project I investigate cell type occurrences in the vicinity of blood vessels in multiplex immunohistochemical stained brain tumour tissue samples. My data includes 9 tissue samples with 18 stain layers each. The entirety of the data is rich, with stain hierarchies corresponding to certain cell types within a fascinating microenvironmental landscape. We select 4 regions-of-interest for each tumour grade per patient 1-3 and map the tumour micro-environment — in particular T cells, macrophage M1/M2, microglia and tumour cells.

We start by registering the nuclei channel of each image and reapplying the transformation to each channel. A pre-trained deep learning model is then utilized in nuclei detection. Next we use nuclei as focal points in watershed segmentation with a mask that is acquired from the blood vessel channel. The subsequent task is composed of measuring mean intensities in filtered stain images of corresponding samples. We acquire an attribute describing distance to nearest blood vessel for each centroid coordinate of our cells — extracted from a feature map based on a processed blood vessel channel.

1 Background

1.1 Digital transformation in healthcare

The digitization of modern civilization has transformed economic sectors; retail, entertainment, finance, education, communication, and so forth — all either fully or partially submerged in the digital space. Whilst healthcare has seen substantial technological development, its adoption is stagnated due to several challenges: the culture isnt set up to absorb innovation, the stakes are very high, regulatory processes persist, lack of clinical insight in technology companies, to name a few examples. [2]

We nonetheless live in a time where healthcare is as relevant as ever: increased chronic disease for an ageing population, health and fitness trends as well as a global health crisis with the Covid-19 pandemic. According to statistics of FiBAN and Finnish Venture Capital Association, the pandemic has created unforeseen demand in digital health, with healthcare being the most attractive sector for Finnish angel investors in 2020 [8].

Sacha Kraus et. al. [9] state that results from a systematic literature review about the current research on the digital transformation of healthcare can be clustered into 5 groups:

- **Patient-centred approaches** Healthcare consumers are becoming active decision-makers in their health management. An article by Patricio et. al finds that participatory service design in healthcare enables innovation. Patient-centric approaches do not replace face-to-face consultation, rather act as a complementing service.
- **Operational efficiencies** Innovations in healthcare technology enhance operational efficiencies, which in turn has a positive correlation with patient satisfaction. These improvements are seen in for example reduction of reaction time and

duration of administrative processes.

- **Organizational factors and managerial implications** Adoption of Health Information Technology is investigated at an organizational level. Organizations can have differing barriers and motivations to digitizing their systems and processes.
- **Workforce practices** Digital transformation in healthcare will reshape the requirements for hospitals and healthcare professionals. These essential requirements include improving on flexibility, development of new behavioural traits and capacity for increased distractions.
- **Socio-economic aspects** For more than a decade, market investments in the European digital health market have been highly associated with socio-economic and technological promises as well as policy initiatives. Socio-economic challenges regarding digitization of healthcare are found within digital literacy, network accessibility, finance and regulatory implications.

Regardless of the stagnation in digitalization of healthcare, quantity of patient-level data has increased to a great extent; 30 % of global data volume is generated in healthcare and the annual data volume growth rate is significantly higher than for manufacturing, financial services and media/entertainment [10]. In addition to volume, healthcare systems possess 4 other distinct *Big Data* related characteristics, as noted in table I [11].

Despite these astronomical data volumes, their potential remain largely untapped. Electronic Medical Records (EMR) were an initial stepping stone to digitization of healthcare, yet they are often attributed to being inefficient and acting as a barrier to innovation [2]. The Finnish Institute of Health and Welfare (THL) states that operating primary and special healthcare as separate entities has resulted in acquisition of several independent systems. This fact combined with private sector having no access to local registries are example challenges in enhancing utility of patient records. The 2023 Finnish reform of social welfare and healthcare will

Table I. The five V's in Healthcare Big Data. Adopted from [11].

Characteristic	**Description**
volume	The large amount of data produced by stakeholders
velocity	The rate at which health data is generated, stored, analyzed, etc.
variability	Changeability in data rate, format, structure, semantics, etc.
variety	Diverse sources of data including multimedia, blogs, server logs, etc.
veracity	The quality of generated data.

bring about the acquisition of unified systems, which will give foundation to further development of electronic services in healthcare. [12]

Variety in data sources is one of the major challenges in digitization of healthcare. These sources include patient demographics, encounters, diagnosis, pathology, laboratory tests, medications, radiology, surgical treatment, post-therapy care and so forth. With the advent of Internet of Things, advanced sensors and mobile technology, additional information is collected. Once a rich data foundation is built on top of these scattered sources, using modern Big Data technologies such as Cloud Services, Parallel Processing and novel Database Management Systems (DBMS), organizations can subsequently enter a new level of operation with next generation data management, advanced analytics, Machine Learning (ML) and Artificial Intelligence (AI). [1, 11]

The aforementioned technologies of advanced analytics and ML/AI[1] are methods related to extracting knowledge from data or other experience as well as generally building intelligent systems. In the case of healthcare this includes for example predicting life expectancy, determining tumour malignancy, segmenting ROI's and

---

[1] AI is a loosely defined umbrella term encapsulating the models that emulate tasks typically requiring human intelligence. ML is a subtype of AI, where learning is from data.

knowledge discovery.

In the experimental part of this paper, I participate in the field of Digital Pathology by investigating the cellular microenvironment of brain tumour tissue samples using advanced image analysis. From the 5 groups of healthcare digitization research mentioned earlier, this work can be classified as belonging to 'operational efficiencies', since understanding and measuring the tumour microenvironment has potential to reduce burden on medical professionals in for example automating tumour grade diagnosis. The next chapters thus accordingly cover some of the background knowledge in digital pathology, the technique used to illuminate the cells (mIHC) and brain tumour microenvironments.

1.2 Digital pathology

The diagnostic discipline on the visual interpretation of captured tissue images, Histopathology, is one of the domains of healthcare whose digitization efforts are being accentuated [13]. Digitization in pathology has been ongoing already for a few decades, although only through recent technological developments have they become propelled. Digital Pathology — the acquisition, management, sharing and interpretation of pathology data in a digital setting — plays a crucial role and is increasingly a technological requirement in the modern laboratory environment [4, 14]. Since the focal point of this thesis is related to advanced image analysis of brain tumour tissue, we focus more on the relevant technologies, implementations, challenges and future prospects of computer vision in Digital Pathology (DP).

Whole Slide Imaging

We can imagine that in order for pathologists to transition to a digital workflow, it'd require a system, in which the primary source of diagnosis — microscopy images — are digitized for display, management, sharing and analysis. This is exactly what Whole Slide Imaging (WSI), a.k.a virtual microscopy, achieves. WSI's are composed of illumination and microscope optical components as well as a system that accurately positions images on a camera — essentially making it a microscope in robotic and computer control. WSI's produce digital renderings of traditional glass slides, which are then viewed with interactive software on a computer. [3, 15]

Although WSI's have been around for more than a decade, they were initially adapted for educational purposes with no clear clinical utility [16]. Recent advancements in processing power, data transfer speeds, software, hardware and cloud services, have deemed WSI as the standard for future, scalable, high-throughput digital pathology [3, 13]. Other digital counter-parts include real-time systems, where microscopes are either robotically controlled from a distance or video-streamed to pathologists.

As mentioned in the previous chapter, healthcare systems are very complex and digitization is not as straight forward as it perhaps may be in other sectors. The technical backbone of digital pathology — big data technologies — require expertise and resources that are not generally found in medical personnel and IT infrastructure. Additionally there can be misalignment between technical and medical goals, with technological implementation lacking clinical insight. Some of the numerous advantages and disadvantages related to WSI's are listed in table II. [13]

Table II. Potential advantages and disadvantages related to implementation of Whole Slide Imaging, in general context as well as in perspective of reporting and user experience. Adopted from [13].

Feature	**Possible advantage**	**Possible disadvantage**
WSI-general	• No physical slide distribution • No slide fading • No lost slides • Faster case finalization • Less misidentification • Easier workload allocation	• Time for evaluable-ready slide increased • Cost of integration • Calibration requirements • Small particles omitted in scan • Image artifacts • Increased IT-dependence
WSI-reporting/ UX	• Parallel viewing • Digital slide superposition • Faster case finalization • Quick access to prior slides • Facilitation of slide presentation • Image sharing • Computational pathology	• Slower evaluation • Interpretation difficulties • Polarization inability • Visibility poor for some structures • Additional training needs • Dual infrastructure (glass & digital)

Computational pathology

Histopathological evaluation of cancer is traditionally performed at multiple scales of magnification and resolution. This is done in order to examine tumour progression relating features, such as nuclei atypia, degree of gland formation, mitosis ubiquity and presence of inflammation. However, human visual assessment is time consuming and incapable of routinely accurately quantifying valuable information such as the hundreds of thousands of cells in every tissue section or mitosis ubiquity [17, 18].

The integration of WSI's to the pathologists workflow — bringing management and sharing of digital slides — opens a new exciting world: the application of advanced algorithms in microscopic images and computer-aided diagnostic techniques in pathology[4]. Histopathological sections contain rich phenotypic information that can be computationally utilized to explore underlying mechanisms contributing to disease progression and survival outcomes [18]; unprecedented development in computer vision allows for tissue interrogation in ways never seen before [16].

Deployment of Machine Learning models and particularly Deep Convolutional Neural Networks are at the center of the new promising potential in practical histopathological diagnostic workflows. Successful applications include tumour classification and segmentation, mutation classification and patient outcome prediction [19]. Next I will cover some of the relevant basics of deep neural networks because of their future prospects and the fact that in my methodology, I utilize a pre-trained deep neural network for nuclei detection.

Convolutional Neural Networks

In machine learning, artificial neural networks (ANN) are a family of models that mimic the structure of the brain and learn patterns in observations. ANN's are composed of computational units that form connected layers. Perceptrons are the most rudimentary ANN's, consisting of an input layer and an output layer, possessing only linear approximation capability. Perceptrons gain non-linear capability by introducing hidden layers between input and output layers, where units of neighbouring layers are connected but units within layers are not. These are called multilayer perceptrons, whose estimation function of an output unit y_k given an input vector $\mathbf{v}$ is as follows: [20]

$$y_k(\mathbf{v};\Theta) = f^2\left(\sum_{j=1}^{M} W_{kj}^{(2)} f^{(1)}\left(\sum_{i=1}^{D} W_{ji}^{(1)} v_i + b_j^{(1)}\right) + b_k^{(2)}\right), \tag{1}$$

where superscript denotes layer index, f non-linear activation functions at specified layers, M number of hidden units and $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}\}$ weight and bias parameter set. There are several possible activation functions, including sigmoid, hyperbolic tangent, softmax and the most popular one – rectified linear unit (ReLu) function [21]. In equation 1, if we regard the hidden layer as a feature extractor $\phi(\mathbf{v}) = [\phi_j(\mathbf{v})] \in \mathbb{R}^M$, the output is a linear model: [20]

$$y_k(\mathbf{v};\Theta) = f^2\left(\sum_{j=1}^{M} W_{kj}^{(2)} \phi_j(\mathbf{v}) + b_k^{(2)}\right), \tag{2}$$

where $\phi_j(\mathbf{v}) = f^{(2)}\left(\sum_{i=1}^{D} W_{ji}^{(1)} v_i + b_j^{(1)}\right)$. The above interpretation also applies when there is increased number of hidden layers, which gives intuition that the role of hidden layers is to extract informative features relative to the target task.

Practically speaking, the model needs to learn parameters Θ that minimize a cost function E. Since the cost function is highly non-linear and non-convex, gradient descent algorithms are used to iteratively update parameters. In feedforward neural networks, the requirement of computing gradients is satisfied with back-propagation of error using the chain rule. Once the gradients are known, parameters Θ can be updated as: [20]

$$\Theta^{(\tau+1)} = \Theta^{(\tau)} - \eta \nabla E\left(\Theta^{(\tau)}\right), \tag{3}$$

where η is the learning rate. There exists several variations to the optimization equation 3 depending on for example how many samples are chosen at each gradient computation and hyper-parameter adaptability [22].

Architectures with multiple layers — deep models — are useful for discovering features from fine to abstract. In histopathology, we are interested in instrinsic cellular patterns exhibited in microscopic images, thus we use models that utilize spatial and configural information in 2D or 3D input, called Convolutional Neural Networks (CNN). CNN's are composed of convolutional layers scattered with pooling and fully connected layers — exploiting mechanisms of local receptive fields, weight sharing and sub-sampling. The units in convolutional layer l compute activation $\mathbf{A}_j^{(l+1)}$ based on a subset of near units in feature map $\mathbf{A}_i^{(l-1)}$ of preceding layer $l-1$ by convolving kernels $k_{ij}^{(l)}$ according to :[20]

$$\mathbf{A}_j^{(l)} = f(\sum_{i=1}^{M^{(l-1)}} \mathbf{A}_i^{(l-1)} * k_{ij}^{(l)} + b_j^{(l)}), \tag{4}$$

where $M^{(l-1)}$ denotes number of feature maps in layer $l-1$, asterisk convolutional operator, $b_j^{(l)}$ bias parameter and f activation function. CNN's require a change in details for gradient descent and backpropagation: the sum of the gradients for a given weight over all kernel weights must be computed — in order to match patches with consecutive layer units and upsample pooling layers to recover reduced sizes. [20]

I have now very briefly introduced neural networks and CNN's whilst covering only the most relevant matter. Deep learning is a vast field with diverse architectures in supervised/unsupervised contexts and rich conceptual framework. Deep learning in medical imaging is used in feature representation learning, detection of anatomical structures, segmentation, computer-aided detection and computer-aided diagnosis. [20]

Standardization of analytics is a hurdle in deployment of Digital Pathology solutions into research and clinical settings. The WSI's we are working with are the product of a series of steps each with contribution to final quality, both in harvesting and processing of tissue. Available scanners have variability and there are image inconsistencies. These protocols and machinery must be standardized in order to deploy reliable DP systems in a larger scale. There is research and solutions in handling of image variability, such as models with generalization capability [16, 23]

1.3 Spatial mapping of brain tumour microenvironment

Brain and other central nervous system (CNS) tumours are constituted by a large collection of histologically distinct subtypes with varying descriptive epidemiology, clinical characteristics, treatments and outcomes. Although these tumours are rare, they compose a high fatality rate and thus a disproportionate effect on cancer mortality; two-thirds of patients die within a 5 year period after diagnosis. In recent years there has been development in the molecular understanding, classification, detection and diagnosis of these tumours, yet the causal factors remain largely unknown and clinical responses limited. [5]

Research has revealed that the tumour microenvironment (TME) plays a major role in tumour progression and response to therapies, which has prompted the urgency to develop methodologies in characterizing these complex tissue sites [24]. Multiplexed immunohistochemistry (mIHC) has emerged as an effective and proficient approach in identifying specific proteins and molecular abnormalities as well as determining spatial mapping and activation of different immune cells [25]. In my methodology, I utilize brain tumour tissue mIHC images courtesy of doctoral student Aliisa Tiihonen from the Cancer Regulation and Immunology research group in Tampere University. In the next section I will take a surface-level look at the mIHC technique used by Tiihonen and components of the brain tumour microenviornment.

Multiplex immunohistochemical staining

Immunostaining is an umbrella term in biochemistry that encompasses all antibody based methods to detect specific proteins in samples [26]. Conventional immunohistochemistry (IHC) is a prevalent diagnostic technique in histopathology, yet it is constrained by high inter-observer variability and capacity to label singular markers per tissue section. These constraints are circumvented by novel so-called multiplexed techniques — particularly multiplex immunohistochemistry/immunofluorescence (mIHC/IF) — permitting simultaneous detection of several markers per tissue sample and subsequent analysis of tissue composition and cell-cell interactions. High-throughput mIHC/IF techniques give a foundation to reproducible, efficient and cost-effective tissue explorations. [7]

In general, immunohistochemistry uses antibodies to detect certain antigens (markers) in a sample of tissue. As antibodies bind to antigens, antibodies can be labelled with fluorophores for detection with fluorescent microscopy or with precipitate-creating enzymes observed by light microscopy. [27] Multiplex strategies of staining multiple antigens with different antibodies grown in same species[2] are essentially conducted either by bleaching directly conjugated primary antibodies before adding layers, blocking access to previous antibody for a consecutive staining round or alternatively removing antibodies from sections after staining and imaging. [28]

The images used in my analysis were produced based on a staining protocol called Multiple Iterative Labeling by Antibody Neodeposition (MILAN). This efficient multiplexing method is implemented with the use of common primary and secondary antibodies, ubiquitous image scanners and routinely processed formalin fixed paraffin embedded (FFPE) tissue sections. MILAN employs the utility of an-

---

[2] mIHC methods can be implemented with anti-bodies from differing species, but these methods are constrained by limited antibody sources.

tibody removal, as mentioned in the last paragraph. [29] The protocol by Tiihonen deviates from MILAN primarily from the fact that she used double staining and conversely did not image autofluorescence. The steps conducted in production of utilized images, have been simplified in table III.

Table III: Main phases in used staining protocol. Process details have been simplified. Various washing steps having been omitted for simplicity. Interpreted from staining protocol document by Tiihonen [30] and steps elaborated by [31].

Step	**Title**	**Description**
1	Tissue setup	Process starts from attachment of tissue sections onto microscopy slides and a incubation period.
2	Paraffin removal	Paraffin is removed from sample by consecutive washes with xylene. Xylene is then removed with graded washes of xylene to ethanol, the sample is then hydrated by graded washes of ethanol to water.
3	Antigen retrieval	Enhance detectability of proteins by putting slides in Tris-HCL -buffer and subsequently to a retrieval device.
4	Autofluorescence reduction	Utilize Sudan Black 0,1 % in incubation of slides in order to block autofluorescence without effecting specific signals.
5	Blocking	Apply normal serum incubation. Carries antibodies that will bind to the non-specific epitopes in sample, thus blocking conjugated antibodies from doing the same.
6	Primary antibody incubation	Dilute the primary antibody in blocking solution and incubate.

7	Secondary antibody dilution	Dilute the secondary antibody in blocking solution and incubate.
8	Nucleus staining	Incubate with DAPI.
9	Slide mounting	Provide a drop of mounting medium on slide and place the the coverslip on top. Remove air bubbles.
10	Imaging	Slides cleaned with Ethanol. Configurations made on imaging device Hamamatsu NanoZoomer s60. Three channels per sample: DAPI, CY5, CY7.
11	Antibody elution	After satisfactory scanning is achieved, coverslip is removed and antibodies are removed. Antibody elution implemented with 2-ME-SDS solution.
12	New staining round	Repeat steps 4 - 11.
13	H&E	After all antigens are stained and imaged, Hemtaoxylin and Eosin staining can be done as final staining.

Tumour microenvironment

Brain tumors are the most common solid tumors affecting children and adolescents — with close to 5,000 children diagnosed each year — rendering them particularly dire [32]. They are notoriously some of the most deadly forms of cancers with ability to resist conventional and novel treatments. In addition to the fact that brain tumours are located inside the most crucial organ of the human body, they are often protected by the blood-brain-barrier (BBB) — a system of tight junctions and proteins that protect neural tissue from factors in general circulation whilst concurrently obstructing the passage of therapeutics. [33]

The tumour microenvironment has been recognized as a key regulator of cancer progression in primary and metastatic brain tumours. It is comprised of several

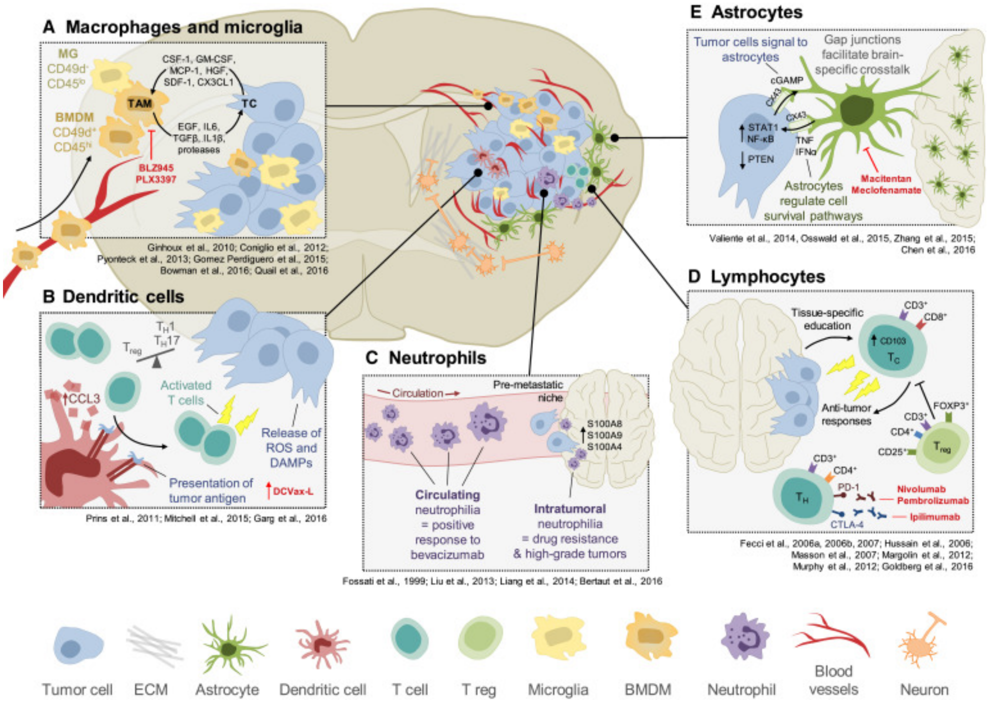

Figure 1. Illustration of the complex landscape that comprises the brain tumour micro-environment. Figure acquired with permission from [6].

non-cancerous cell types in addition to cancer cells, all interconnected in a highly complex landscape, as illustrated by figure 1. Cellular components of the brain TME include tumour cells, various immune cells and organ residents such as neurons and astrocytes, listed in table IV.

Historically, the brain was thought to be one of the 'immune privileged' organs in the body due to to several reasons. However this perspective has shifted with discovery of things such as functional lymphatic vessels in dural sinuses of mice, CNS-derived immune responses and compromised BBB's that permit entry for multiple types of immune cells from peripheral circulation. There are several intrinsic interactions within the TME that are associated to tumour progression, and are thus under investigation for potential therapy. In this regard, I will introduce a few relevant features in the following. [6]

In removed glioblastoma tissue, tumour-associated macrophages (TAMs) and

Table IV. Cellular components present in the brain TME [6]. Definitions from [27].

Cellular component	**Description**
Cancer cell	Malignant cells constituting the tumour
Extracellular matrix	Network of molecules that surround, support and give structure to cells and tissues.
Astrocyte	Glial cells that hold nerve cells in place and develop as well as maintain their functionality.
Dendritic cells	Immune cell that boosts immune responses by showing antigens on its surface to other immune cells.
T lymphocyte	A type of white blood cell that develops from stems cells in bonemarrow. Protects the body from infection.
Regulatory T cell	A type of white blood cell that regulates immune system activity.
Microglia	Type of glial cell that acts as one of the two main immune defenses in the CNS.
BMDM	Bone marrow-derived macrophages. A type of white blood cell that surrounds and kills micro-organisms, removes dead cells and stimulates action of other immune cells.
Neutrophil	A type of white blood cell that ingest and kill microorganisms at site of infection. Additonally boosts response of other immune cells.
Neurons	A type of cell that receives and sends messages from the body to the brain and vice versa.

organ-resident microglia reportedly constitute up to 30 % of tumour mass. TAMs are derived from the bone marrow through the action of glioma and other cells. Studies have shown that glioma cells suppress immune surveillance of TAMs by altering them towards pro-tumourigenic M2 phenotypes while inhibiting development of anti-tumourigenic M1 phenotype. It has also been observed that glioma cells can induce a mixed population of M1 and M2-related molecules. It can nonetheless be stated that glioma cells and TAMs have complex bidirectional communication, with TAMs supporting tumour growth and progression. Using immunohistochemistry, it has been shown that number of M2-related markers had a positive correlation with tumour grade. It is however important to note that the M1/M2 dichotomy has become obsolete; many groups focus on defining context-specific macrophage activation and phenotype as a measure of functional diversity. [34]

T cells are released from the bone marrow and thymus into the peripheral system, from which they expand to other tissue after they have matured. T cells are part of the adaptive immune system, meaning that they are activated on the surface of antigen-presenting cells, which mature during the innate immune response triggered by an infection. In human glioma samples, a study found that the brain TME altered T cell effector functions during anti-tumour immunity, despite being in an environment where T cell responses are minimal under normal conditions. T cell activation through the use of inhibitors is a emerging field of cancer therapy. [6]

Other distinct features of interest include prognostic value of neutrophils, regulation of BBB from TME components and the roles of organ residents such as neurons and astrocytes in tumour initiation and progress [6]. Generally speaking, researchers are interested in immunotherapeutic treatment for brain cancer, thus rendering the TME vital to understand. It is apparent that there is a plethora of research to conduct, with the brain requiring a specific framework for designing TME-targeted therapies. In this MSc thesis in analysis of mIHC tissue samples of glioma, we are particularly interested in macrophage and T cell populations in the vicinity of blood vessels for different tumour grades.

1.4 Literature review

In order to conduct spatial analysis of the tumour environment — whether it be studying immune cell proportions or anything else — it is necessary to computationally map the tumour microenvironment so that cell types are recognized as distinct class labels. This is a complex and laborious task, which is indeed the foundation of our problem scenario; there is a plethora of interesting questions to answer regarding spatial properties of the tissue, after the expression of the receptors of interest have been discretized.

Danielle J. Fassler et al. [35] create an ensemble method *ColorAE U-NET* from autoencoder *ColorAE* and convolutional neural network *U-Net*. *ColorAE* predicts an 8-channel image of the same size of the input image. The channels correspond to the concentration maps of six IHC stains in addition to hematoxylin and background channel. Their ground truth data included a set of 80 patches from 10 mIHC WSI's and utilized seed labels as well as super pixels to gain a relatively large training set. A pathologist examined each patch and placed a seed annotation to each cell in order to indicate cell type based on staining. In training, a reconstruction loss compares reconstructed image to input-image using mean-squared-error. Furthermore, they leveraged weak-form supervision from human annotators through a label consistency loss-function. The second model was created based on U-Net architecture — it was trained to generate features that differentiate cell classes according to provided labels. The results from ensemble method *ColorAE U-NET* were used in a subsequent proof-of-concept spatial analysis to calculate average distance between cell types.

Yeman Brhane Hagos et. al. [36] propose *ConCORDe-Net* — a cell count regularized convolutional neural network for cell detection and classification in mIHC images. Inspired by *inception-v3*, their model integrates conventional dice overlap and a novel cell count loss function for optimizing cell count, followed by multi-stage CNN for cell type classification. Their training data consisted of over twenty-thousand annotated cells, belonging to five classes annotated by experts from 175 patches extracted from 6 tumour WSI's. Using *ConCORDe-Net*, they obtained an F1-score of 0.873 for cell detection. The model excels at detecting densely packed cells and weak-staining compared to other state-of-the-art methods. Furthermore, a 96,5 % cell type classification accuracy was reported.

Yanan Wang et. al. [37] built *Cell Graph Signature* — a novel graph neural network-based approach for prediction of patient survival in gastric cancer. Their cohort consisted of 172 gastric cancer patients with records on their survival status. Two data binning strategies were applied in order to achieve binary- and ternary-class datasets in terms of survival. In their pre-processing of mIHC images of 7 biomarkers, each image was first segmented into multiple non-overlapping regions. For each region, a graph was built with each cell type represented as a node and edges defined as reciprocal of distance between cells. Subsequently 35 features were extracted for each node resulting in a so-called Cell-Graph dataset. 90 Cell-Graphs were constructed for each mIHC image for each patient. Using cross-validation and hyper-parameter optimization they tested 4 different models with *GINTopK* having the best results and selected as final model. They reported an astounding AUROC value of $0{,}96 \pm 0{,}01$ on five-fold cross validation for the binary class prediction.

2 Problem setting

2.1 Problem statement

Despite success in treatment of several types of cancer, brain tumours remain especially devastating with a high mortality rate and relative prevalence in children/adolescent. The understanding of the brain tumour microenvironment has emerged as critical in regards to cancer progression and potential therapeutic solutions. Tumour microenvironments are efficiently mapped with novel multiplex immunohistochemistry staining methods. Due to the big data nature of histopathology, only through recent technological developments has its digitization become generally feasible and increasingly a superior system to traditional methods. With the capacity to map tumour environments and manage the staining output digitally, we can then seek to use image analysis in hopes of automating tasks to reduce burden on medical professionals, improving clinical responses and knowledge discovery.

2.2 Objectives

In this Masters thesis I use computational methods in mapping spatial distribution of immune cell populations within diffuse glioma mIHC tissue samples and subsequently analyze cell type occurrences in the vicinity of blood vessels. I aim to create a program, that from a given image, extracts structured data of cells with information on their location, type and distance to neareast blood vessel. Once data is formed in the desired format, the subsequent analysis of relative cell quantities and spatial qualities, is enabled. In order to conduct spatial analysis, the following primary steps are implemented:

1. **Image registration** The stainings for each sample are registered using a image registration module based on affine transformations

2. **Nuclei detection** Pretrained deep convolutional neural network StarDist2D is applied in nuclei labelling in DAPI images

3. **Cell segmentation** Nuclei labels are given as input for watershed algorithm, which hypothetically approximates cellular area

4. **Feature extraction** Calculate intensity properties from cell areas of thresholded stain images and modify data to desired format

2.3 Scope and Limitation of the Study

The purpose of this thesis is to investigate immune cell proportions and spatial distributions in the vicinity of blood vessels in diffuse glioma mIHC sample region-of-interests. We compare the aforementioned properties between ROI's of different samples, since they have different grade tumours and differing immune responses. The spatial mapping of the tumour microenvironment is a pre-requisite to our analysis, and is thus under focus. We also highlight the utility of developing computational methods for multiplex immunohistochemistry outputs.

The mapping of cell types in mIHC is a complex problem: TME consists of a plethora of components of varying structure, antigen receptors are not definitive, the images are Gigabyte-size, and so forth. At the core, we are trying to map intensity distributions to distinct labelled cells on an image. Whilst it is possible to use traditional methods in segmenting the cells and calculating intensity vectors — it comes with a lot of constraints. Each image section must be processed with multiple algorithms and the program output data is modified to cell-wise. There are hundreds of thousands of cells per sample with morphological attributes that are out of the scope of this project.

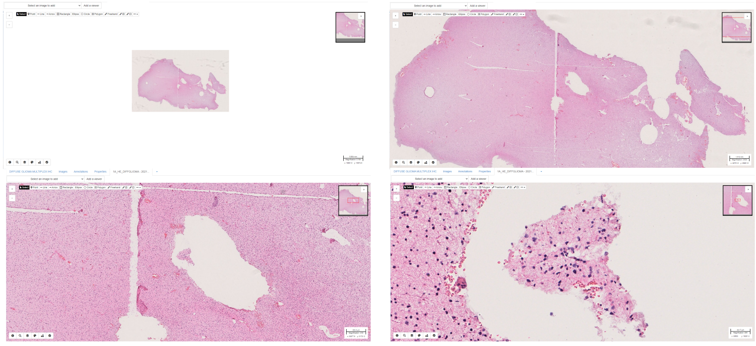

Figure 2. Hematoxylin & Eosin stained mIHC tumour tissue at different levels of magnification viewed in Cytomine. The pink and black illumination signify extracellular matrix and cell nuclei respectively.

The program is suitable for densely packed tissue areas; utilized segmentation of cells is not reliable in spatially inhomogeneous regions-of-interest. Using traditional methods, structures with no visible nuclei such as blood vessels require special attention. Ultimately, the task is solved best with deep learning — this approach was not taken due to lack of labelled data for target task and time constraints. Training a deep learning model for this approach would optimally require hundreds or thousands of labelled ground truths for desired segmented cell types.

3 Methodology

3.1 Image acquisition

Cytomine is an open source web application for collaborative analysis of multi-gigapixel images designed for development of machine learning, image informatics and big data. As a Cytomine user, you have capability to upload large images characteristic to histopathology, view images at different magnifications, manage/configure projects and add various annotations associated with terms. [38]

Aliisa Tiihonen conducted mIHC over a period of approximately a month for 4 patient cases resulting in an impressive 18 antigens used per sample — as noted in table V — in addition to measuring H&E and separate cell nuclei illuminating DAPI channel. After completion, she uploaded all the images to Cytomine into the project folder. The images came in different sizes with pixel width $p_x \in [50176, 69888]$ and pixel height $p_y \in [37632, 53760]$ as well as 20x magnification. All stain images excluding H&E were transferred to compute cluster Narvi as presented in next section.

Case information was provided by Tiihonen, as illustrated in table VI. In our case definition — e.g. 1A — the number signifies the patient and the letter denotes tumour progression. For example case 1A is tissue from the first diagnosis (III diffuse astrocytoma) for a 32 year old female patient and 1B is tissue, that is collected 7 years later from the same patient (relapse). Note that for patient 2, the order of samples is illogical — with 2C being the primary diagnosis, 2A first relapse and 2B the second relapse. Additionally, information on survival in months after glioblastoma (GBM) diagnosis is reported.

Table V. Antigens measured by Tiihonen and their corresponding receptors. All images were registered, but subset colored in gray was selected for further analysis.

Antigen	**Receptor**
CD45	All blood cells except mature red blood cells and platelets
CD3	T cells
CD4	Helper T cell
CD8	Cytotoxic T cell
FOXP3	Regulatory T cell
CD20	B-lymphocytes
CD56	NK-cells
CD66B	Granulocytes
TMEM119	Microglia
CD68	M1-macrophage
CD163	M2-macrophage
CD11c	Dendritic cells
CD31	Blood vessels
NeuN	Neurons
IDH1	Tumour cells
Ki67	Any cell, proliferation
MHCI	Any cell, antigen ability
pSTAT3	Any cell, poor prognosis
H&E	ECM, Nuclei (Tissue morphology)

Table VI. Patient case information.

Case	**Gender**	**Age at diagnosis**	**Primary Grade**	**Survival after GBM diagnosis [months]**
1A	Female	32	III diffuse astrocytoma	
1B	Female	39	IV glioblastoma	13
2C	Male	29	II diffuse astrocytoma	
2A	Male	34	III diffuse astrocytoma	
2B	Male	41	IV glioblastoma	17
3A	Male	30	III diffuse astrocytoma	
3B	Male	31	IV glioblastoma	25+
4A	Female	41	III diffuse oligodendroglioma	
4B	Female	44	IV glioblastoma	9

3.2 Computational environment

Access to compute cluster Narvi was admitted by Tampere University computing centre (TCSC) and supervised by PhD student Masi Valkonen. Narvi is built on SLURM — a cluster management and job scheduling system for Linux clusters — boasting 140 CPU-only nodes with 3000+ cores and 22 Tesla GPU nodes (V100/P100/K80) with 4 cores each. SLURMs functionality as a workload manager includes resource access management, framework for starting, executing and monitoring work as well as queue management. [39, 40]

A Python virtual environment was created within the Narvi project folder using Anaconda. Programming was implemented in Python and bash scripts with testing done in Jupyter Notebooks. Inconvenient scanner output names needed to be renamed according to information from Tiihonen on channel output pairs. Hierarchical relationships in stains were modelled as a basic class with functionality to fetch stains for desired receptors and output their paths for a given sample.

3.3 Image registration

In order to being able to measure stain intensities for a given area, the images need to be in the same coordinates. Image registration was implemented with a module provided by Masi Valkonen. The registration was based on affine transformations, which consist of two primary steps [41]:

1. **Determine the corresponding points between the images**
    - Use labelling functionality in Cytomine
    - Gather 3 corresponding points r $\in [1, 2, 3]$ in all stain images

2. **Determine the transformation between the corresponding points**
   - Solve affine transformation matrix T from equation 5 relative to sample DAPI
   - Reapply transform to images for registered output

Affine transformations

The affine transformation (AT) matrix **T** consists of rotation $\mathrm{T_r}$, translation $\mathrm{T_d}$ and scaling / gain $\mathrm{T_s}$ components, which are represented by parameters a_{ij} in the following AT equation [41]:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \iff \mathbf{p}' = \mathbf{T}\mathbf{p}, \tag{5}$$

which is a generalization of linear transformation and where x', y', z' are target image coordinates, x, y, z corresponding coordinates of image to be transformed and the right hand the vector form equivalence. From equation 5 we can state:

$$x_i' = a_{11}x_i + a_{12}y_i + a_{13} \tag{6}$$

$$y_i' = a_{21}x_i + a_{22}y_i + a_{23}. \tag{7}$$

These two sets of linear equations are of the form:

$$\mathbf{M}\mathbf{a} = \mathbf{b}, \tag{8}$$

which have a least squared solution:

$$\mathbf{a} = \left(\mathbf{M}^{\intercal}\mathbf{M}\right)^{-1}\mathbf{M}^{\intercal}\mathbf{b}. \tag{9}$$

During the staining process, DAPI was measured for each double staining round. In the registration program, equation 9 is used to transform all DAPI's to the same coordinate system — the DAPI from CD45-IDH1 is kept as the fixed image whilst the rest as moving images. The solved transformations parameters are then reapplied to each corresponding stain and subsequent *bicubic interpolation* is performed to interpolate pixels on the grid.

3.4 Artefact removal

Cell detection and classification are particularly challenging for mIHC images due to high levels of variability in staining, expression intensities and inherent noise as a result of preprocessing artefacts. Generally in IHC, common artefacts include non-specific binding, high background noise, overstaining and weak-staining. High background noise and overstaining can result in important features and tissue structures being obscured by the background signal and thus being indistinguishable. Conversely, weak-staining results in very faint signals. Non-specific binding — the binding of primary or secondary antibodies to something else than the designated target — is particularly difficult and results in false positives. [35, 42]

In our dataset, there is high variability in expression intensities and background noise. Activation thresholds for each selected antigen were defined by looking at heat maps of each channel and determining suitable values based on expert knowledge from Tiihonen. As an example, the activation of CD163 and its corresponding filtered image is illustrated in figure 3. For the sake of clarity of color scale, values above 10 are taken as 0, resulting in dark areas in local maxima. In our program, we simply apply a threshold to each image, as determined by table VII.

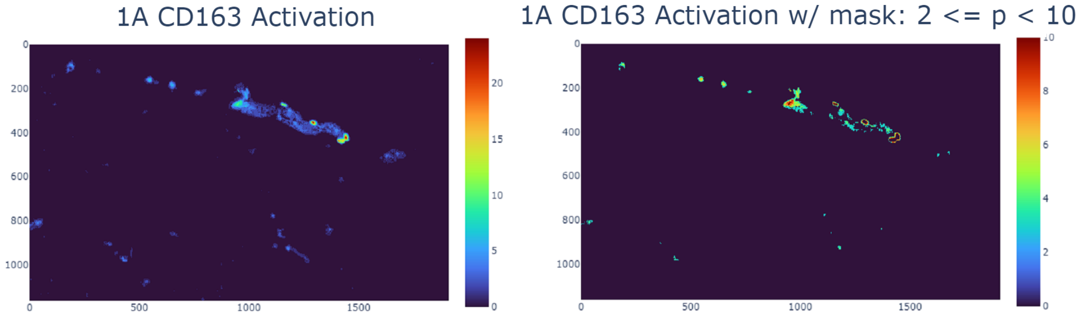

Figure 3. CD163 raw signal and corresponding filtered image of local area in sample 1A.

Table VII. Utilized threshold values for selected antigens of interest.

Antigen	**Threshold**
CD45	3
CD3	3
TMEM119	2
CD68	2
CD163	2
CD31	1
IDH1	6

3.5 Nuclei detection

StarDist 2D is a pretrained convolutional neural network that facilitates automatic detection and segmentation of cell nuclei. Whilst there are other successful learning-based methods for cell nuclei segmentation, StarDist performs well in situations of crowded cells — making it fitting to the densely packed landscape seen in brain tumour tissue. It localizes cell nuclei via star-convex polygons, which are better shape representations compared to bounding boxes. For each pixel, distances to the boundary of the object to which the pixel belongs to $d_{i,j}^{k}$ are regressed, along a set of n predefined radial directions with equidistant angles. Additionally, the model separately predicts whether a pixel is part of an object and subsequently *non-maximum supression* (NMS) is performed to these polygon-proposals and their associated probabilities $d_{i,j}$. [43]

StarDist has been implemented with basis in U-Net — the de facto biomedical image segmentation CNN. U-Net architecture consists of a contracting and expansive path, as illustrated in figure 4 on the left and right side respectively. The contracting path applies repeated two unpadded 3x3 convolutions with a subsequent ReLU activation and 2x2 max pooling with stride 2 for downsampling. At each downsampling step the number of feature channels is doubled. The expansive path consists of layers of upsampling feature map followed by 2x2 convolution that halves number of channels, concatenation with corresponding cropped contracting path and two 3x3 convolutions with subsequent ReLU activations. At the final layer a 1x1 convolution maps each 64-component feature vector to the desired number of classes. High resolution features from the contracting path combined with upsampled output and the large number of feature channels in upsampling, respectively provide localization capability and context information propagation to higher resolution layers. [44]

StarDist is composed of an additional 3x3 convolutional layer with 128 channels and ReLU activation after the final U-Net feature layer. For the object probability

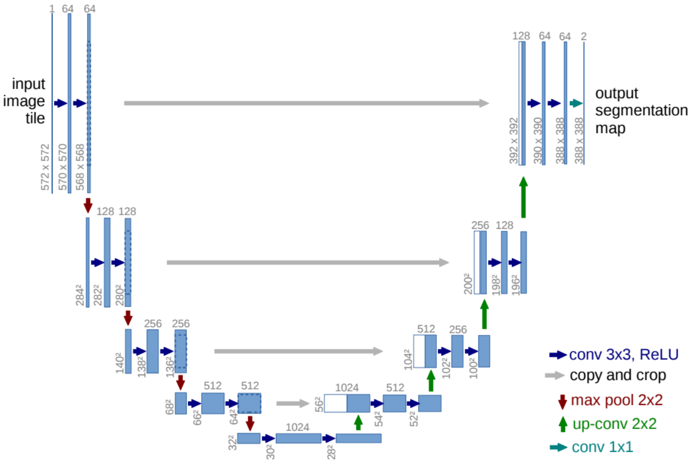

Figure 4. U-Net architecture. Blue boxes signify multi-channel feature map. Number of channels and image size are denoted on top of and on left hand side of boxes respectively. Significance of each arrow is represented in the right side of the image. Figure acquired with permission from [44].

predictions, a single-channel convolutional layer with *sigmoid* activation is used. The polygon distance output layer has same number of channels as there are radial directions n and no activation. In training, *binary cross-entropy loss* is used for predicted object probabilities and *mean absolute error* weighted by ground truth object probabilities for polygon distances. StarDist was tested and compared with U-Net variations on various popular cell nuclei datasets: the average precision for StarDist was superior with smaller *intersection over union* thresholds — as expected for predicting a parametric shape model instead of per-pixel basis. [43]

In this thesis project, the goal is essentially to detect location of different immune cells in brain tumour samples. Since each cell contains a single cell nucleus, predicted instances from the StarDist 2D model can be used for further analysis. The application of StarDist 2D to two different sized images, $(204, 124)$ and $(1912, 1164)$, from sample 1A is presented in figure 5. The top-center image shows the star-convex shapes with 16 radial directions $d_{i,j}$ predicted by the model. Each predicted instance is then associated with a label, as illustrated in the right side column of the

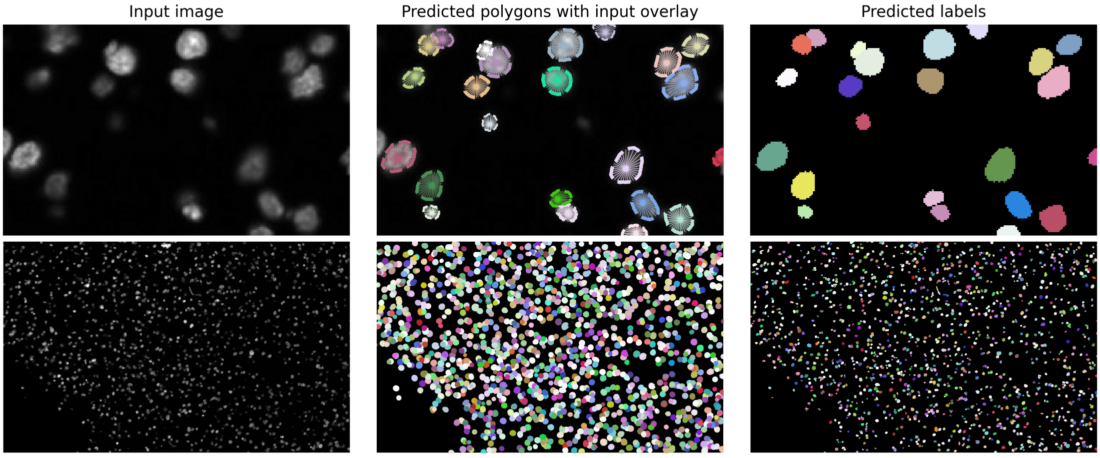

Figure 5. Pretrained model StarDist 2D applied to two different sized input images (left). Predicted star-convex polygons presented in middle; 16 radial directions visible in smaller input (center top). Corresponding labelled instances on right side.

image. The model was tested in various subsets of sample images with qualitatively successful results.

3.6 Cell segmentation

Watershed by flooding is a prevalent segmentation technique for gray-scale images, which characteristically simulates a flooding process. Images are represented as a topographic relief, where the altitude of each point is determined by the gray-scale value of its corresponding pixel. Subsequently, holes are pierced in all regional minima within the relief. The whole surface is then slowly sunk — water springs through the holes and progressively immerses the adjacent walls. An obstruct is set at the conjunctions of the basins and as the relief is fully submerged, the set of obstructs depict the segmented image. Watershed can also implemented by so-called *rain falling*, that uses connectivity components and models the rain process. [45]

Watershed by flooding consists of the following steps [46]:

1. Identify local minima and assign each one with a unique label

2. Simulate flooding process with a priority queue consisting of H queues, where H is the number of possible element values e.g. pixel range. When the queue is full, image elements with value h are pushed to corresponding queue with number h.
    - Fill priority queue with elements of local minima
    - Scan priority queue in sequence from smaller to larger values of h
    - Select element from first non-empty queue — If all queues are empty, terminate algorithm
    - Remove selected elements from queue and propagate marker to all unmarked neighbours
    - Place all marked neighbours from previous step into queue, repeat step 2

In this thesis project, we use StarDist 2D output labels as input for the watershed algorithm provided by the open source Python library *scikit-image*. First the DAPI image is converted to gray-scale. Subsequently we apply the the distance transform, in order to improve segmentation results. Distance transforms label each pixel of the image with the distance to the nearest boundary pixel e.g. the background distinguished in gray-scale transformations. Taking the inverse of distance transform results in cell nuclei locations depicted as local minima. Applying watershed to the inverse distance transform essentially floods the image from each cell nuclei and draws borders at the point of meeting, as illustrated in figure 6. Each hypothetical cellular area is now associated with a distinct label.

Structures such as blood vessels require special attention since they are not associated with nuclei and their corresponding antigen CD31 does not clearly mark the boundaries of the structures. In the DAPI input image of figure 6, the outlines of a blood vessel is visible as a worm-like non-nuclei area in the center top. It is apparent that applying watershed without further specification results in a segmentation error regarding blood vessels. We solve this problem by acquiring a mask from the CD31 channel. The signal from CD31 is weak and thus we apply morphological operations

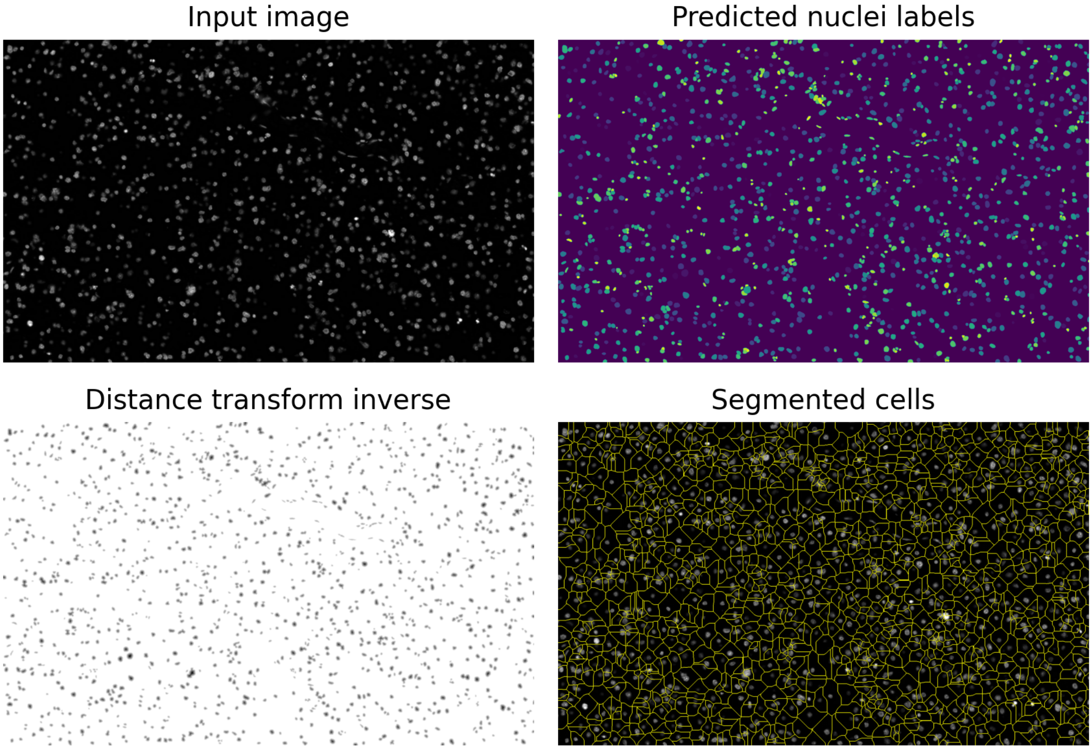

Figure 6. Cell segmentation process from left to right: 1. Input DAPI image 2. StarDist prediction cell nuclei labels 3. Distance transform inverse of gray-scale StarDist prediction 4. Segmented cells in Watershed algorithm output.

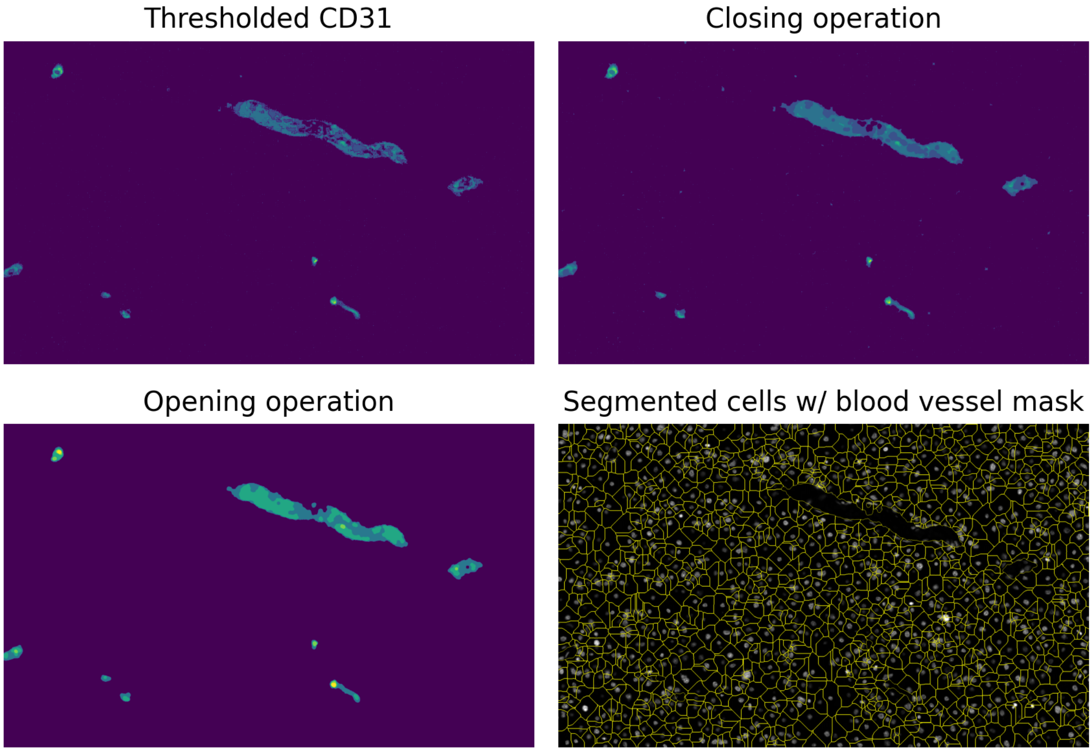

Figure 7. In order to exclude blood vessels from cell segmentaion, a mask is acquired from processing CD31. Top left: CD31 channel with thresholded activation. Top right: Closing operation applied to thresholded image. Lower left: Opening operation applied to closed image. Lower right: Watershed segmentation with blood vessel mask.

— closing with subsequent opening — in order to improve outlines and fill in empty space in blood vessels. In our watershed function call, we set the mask as such, that it is true else where besides blood vessel associated pixels. This process is illustrated in figure 7, with blood vessels now being excluded from cell segmentation. As a final step, we apply a function that removes regions with minimal or maximal pixel-wise areas as well as those that have high perimeters.

3.7 Feature engineering

In order to acquire structured data from the cell-segmented image, we utilize a region property extractor. We define a function that takes segmented cell labels and a list of all corresponding stain channel images of our ROI as input. This function computes intensity properties — min, max and mean — as well as centroid coordinates for each cellular region in each stain image. The output of this function results in a 2D table with each region measurement as its own sample. Since there are several measurements per region, we apply pivoting for each intensity property against our stain labels, resulting in cellular regions as rows, and intensity properties for each channel as additional columns. Both formats — wide and long — are interchangeably used in our analysis according to whichever is suitable for a given task.

Since we are interested in cell types in the vicinity of blood vessels, we need a measurement for this scenario. To solve this problem, we acquire a feature map from the opened image of figure 7. The background and foreground of blood vessels are easily distinguishable after the opening operation so we create a new masked image with the blood vessels as 0 and the rest as maximum value. Now, applying the distance transform to this inverted image results in a feature map as illustrated in 8, where each pixel is defined by the distance to the nearest blood vessel boundary. For each one of our samples we acquire the corresponding blood vessel distance feature with the centroid coordinates.

In defining stain activation for a given segmented cell, we are faced with a rather arbitrary task of choosing an activation threshold. We define a cell sample as activated for a given stain, if atleast 5 % of the cell area is above the threshold in that channel. In figure 9, we have illustrated the relationship between proportions of labelled cell types and the activation threshold used in defining them — for a given extracted dataset from sample 1A ROI. The proportion of unidentified cells

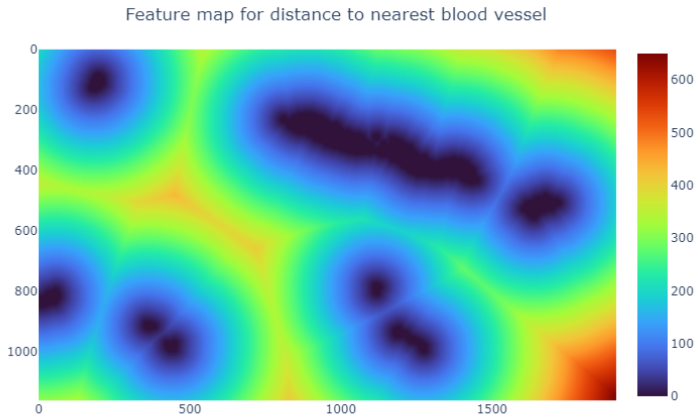

Figure 8. Feature map acquired from CD31 channel. Pixels are defined by distance to nearest blood vessel boundary.

starts off at about 10 % when activation threshold is defined as $t = 0{,}01$ and steadily increases to 60 % when $t = 0{,}5$. Given the variability of multiplex staining outputs and the overestimation of cellular boundaries, it is difficult to evaluate a universally suitable activation threshold.

We are also faced with an additional issue: given cellular regions can have activation for more than one stain, especially with natural hierarchies being present — for example CD45 is activated for all immune cells and CD3 is a child of CD45 that is activated for all T cells. Furthermore, T cell subtypes have their own associated specific activation, such as CD4 for helper T cells. We however, are investigating T cells as a general group, as signified in table V. Double activations that we investigate are microglia (CD45, TMEM119), T cells (CD45, CD3) and M1-macrophages (CD45, CD68). M2-macrophages are the sole triple activation we are taking into account (CD45, CD68, CD163). In addition, we create a new stain group for those samples with only CD45 activation. These are considered to be ’other immune cells’. We also measure the number of samples with no activation.

It is important to note that there is uncertainty that cell types must have all cor-

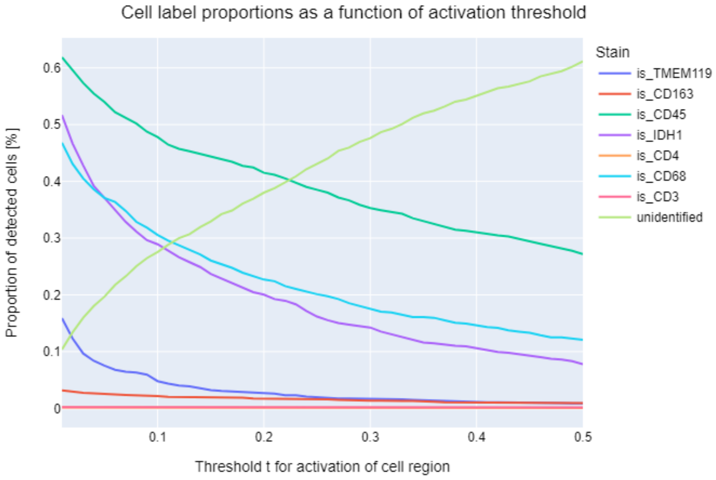

Figure 9. Proportion of labeled stain activations plotted against region activation threshold for given stain. Data extracted from same ROI as in figure 7, with resulting 1397 total cells. Note that in this graph, samples can have multiple activations, for example CD45 being present in all immune cells. This is apparent in the graph, as taking the sum of proportions for a given threshold does not equate to 1.

responding hierarchical stain activations, in for example M2-macrophage requiring all 3 activations instead of solely CD163 — thus we calculate activations of all aforementioned combinations and compare proportions. Additionally, certain regions can have activations for multiple stains that do not pose any real biological significance. This can be a result of mIHC artefacts, imprecise image registration or inaccurate cell regions.

Pseudocode for utilized data extraction program is presented in **Algorithm 1**. It essentially combines everything we have described thus far: for the given sample, biomarkers and corresponding stains — nuclei are detected, watershed segmentation with blood vessel mask is applied and region properties are extracted as well as processed and saved. The program output of our data extraction program — with intensity properties, corresponding defined cell labels, centroid coordinates and distance to nearest blood vessel — is used in spatial analysis.

Algorithm 1 Pseudocode for utilized data extraction program for mIHC.

```
procedure MAIN
    biomarkers ← list of biomarkers of interest                ▷ e.g. immune cells
    stainLabels ← getStains(biomarkers)                        ▷ corresponding stains
    sample ← desired sample                                    ▷ e.g. 1A
    stainPaths ← getStainPaths(sample, stainLabels)
    dapiPath ← getDapiPath(sample)
    roiCoords ← getROIcoords()
    thresholds ← HashMap                    ▷ key-value pairs for noise thresholds
    stainImages ← empty list
    bloodVesselMask ← empty
    for imagePath, stain in stainPaths, stainLabels do
        image ← imagePath.open()
        image ← image.crop(roiCoords)
        artefactMask ← image $\leq$ thresholds[stain]
        if stain == 'CD31' then
            bloodVesselMask ← createBloodVesselMask(image)
        end if
        image[artefactMask] = 0
        stainImages.append(image)
    end for
    dapi ← dapiPath.open().crop(roiCoords).rgb2gray()
    cellLabels ← segmentDapiCells(dapi, bloodVesselMask)
    cellLabels ← filterCellLabels(cellLabels, dapi)
    regionProps ← multichannelRegionProps(cellLabels, stainImages, stainLabels)
    featureMatrix ← processRegionProps(regionProps)
    featureMatrix ← addDistanceFeature(featureMatrix)
    featureMatrix.toCSV()
end procedure
```

3.8 Analysis

We add an additional function to our data extraction program, in which we calculate stain activation absolute and relative proportions as well as mean distance to nearest blood vessel by stain group. We select 4 ROI's of size $p_x = 3822$ and $p_y = 2330$ from each tumour grade from samples 1 - 3. We define the data extraction program to support efficient parallelization on a SLURM cluster. Each program can now be run as tasks in separate nodes within the cluster — computation is simultaneous for all ROI's. The program execution time per ROI is approximately 300 - 500s depending on DAPI image size. In the output we get all cases with extracted CSV's, excel file for analytical properties, and figures related to segmentation as well as ROI mapping. We create an additional script, that combines all the proportion and mean distance to nearest blood vessel data into one dataframe.

Essentially, most of the per-ROI information provided by the full data of our data extraction program is illustrated in figure 10. In this graph, sample location is plotted according to centroid coordinates, coloured by determined cell label and sized by distance to nearest blood vessel. This plot gives us validation — that the blood vessel distance feature has been correctly parsed — since the cells near empty sites (blood vessels) are indeed small, as required. This image was produced for observational purposes — it is smaller in size than actual ROI's, multi-activations have been omitted and cell labelling done ad-hoc — this image is not part of our results.

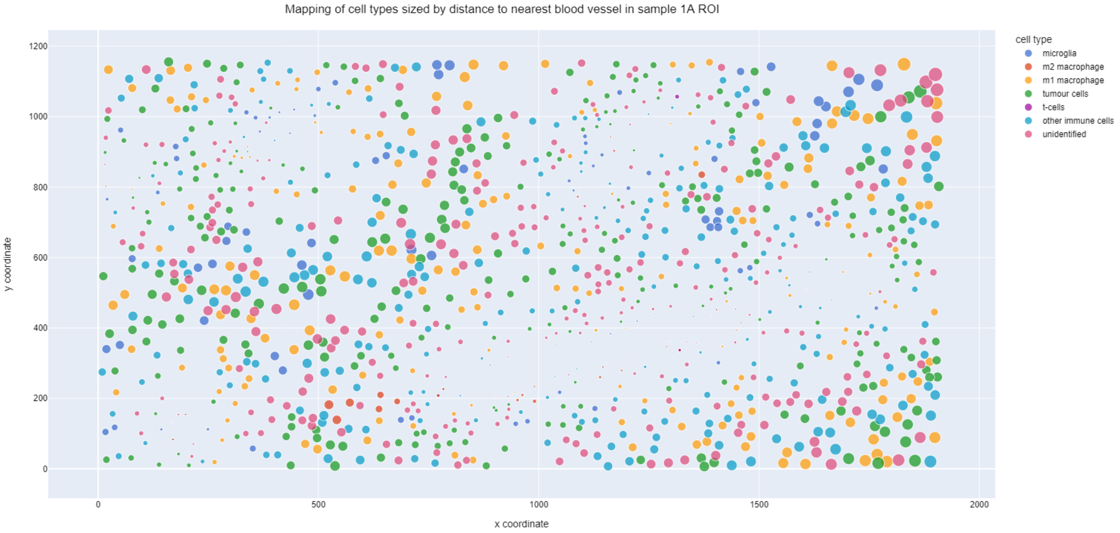

Figure 10. The mapping of cell types in same ROI as in 7. Cells are marked on the map based on their centroid coordinates, colored by their cell label and sized by their distance to nearest blood vessel. The image and its constituent cell labelling were produced for observational purposes and are not part of our results.

4 Results

4.1 Results

Results from all ROI's per sample have been listed in tables VIII - XIV. We've included the activation occurrence N, relative proportion $N[\%]$ and mean distance to nearest blood vessel per activation group $\overline{b}_{stain}$. Additionally the mean of activation proportion and distance to nearest blood vessels from all ROI's per patient per sample are illustrated in figures 12 - 14. We analyze the results at sample and patient level. The activation occurrence proportion from all groups in 1A ROI's are relatively consistent. As expected, immune cell, macrophage, microglia and tumour cells are in abundance — and specialized T cells are low. The mean distance to nearest blood vessel is small for CD163 and CD3 activations, implying that T cells and M2-macrophages are near blood vessels at this point in time for this tumour grade. It is important to note that all immune cells travel to the brain through blood vessels despite the only organ resident immune cell type, microglia.

Looking through the results, most of the segmentations worked well with similar appearance to figure 11. The biggest challenge in terms of our results is staining variability. This is especially true for sample 1B, with inhomogeneous tissue regions and high background noise. We applied an increased CD31 threshold for this sample, in order to acquire a desirable segmentation result. We can nonetheless observe that CD3 and CD163 are substantially higher in occurrence in relapse sample 1B grade IV glioblastoma compared to sample 1A grade III astrocytoma 7 years prior. It is also noteworthy that mean distance to blood vessel has stabilised to similar values of other stain groups, perhaps implying that in the grade IV tumour, T cells and M2-macrophage are more homogeneously distributed.

Sample 2 is the most interesting, since it has data for three grades: II/III (C/A) astrocytoma and IV glioblastoma (B). 2C (lowest grade tumour for patient 2) occurrence levels are comparable to 1A, with less activation for M2-macrophage and interestingly more for microglia. 2A has similar proportions as 2C, with CD45 — and thus general immune response — being on average higher in 2A. In 2B the differences are more visible: M2-macrophage and T cell occurrence levels as well as mean distance to blood vessel being on average higher. Additionally, microglia expression is less in 2B compared to 2A. Sample 3 has large quantities of M2-macrophage in both 3A and 3B, with it being greater in the latter. 3B has a relatively high amount of T cells compared to 3A. We make the same observation as in sample 1, in that mean distance to nearest blood vessel seems to grow for M2-macrophage and T cells with increasing tumour grade.

	N	N [%]	$\bar{b}_{stain}$	N	N [%]	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	121	2.2	49.40	117	2.14	35.85
CD3	5	0.09	54.86	3	0.05	64.13
CD45	2984	53.20	236.79	2600	47.67	172.32
CD45CD3	3	0.05	20.8		0	-
CD45CD68	1322	23.60	234.92	996	18.2	2
CD45CD68CD163	69	1.23	59.04	42	0.77	40.42
CD45TMEM119	307	5.48	46643	331	6.06	150.22
CD68	1992	35.56	233.74	1640	30.06	164.75
IDH1	2496	44.56	246.53	3356	61.53	172.49
TMEM119	507	9.10	210.01	557	10.21	142.60
Only CD45	871	15.55	236.15	576	10.56	165.22
No act.	1009	18.01	225.14	788	14.44	151.90
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	103	2.22	47.69	79	2.34	47.81
CD3	2	0.04	60.34	3	0.88	116.83
CD45	1784	38.50	175.03	1397	41.36	225.3
CD45CD3	1	0.02	11.40	3	0.88	116.83
CD45CD68	569	12.28	165.47	625	18.50	221.73
CD45CD68CD163	37	0.80	44.01	29	0.86	51.05
CD45TMEM119	247	5.33	147.98	115	3.41	136.95
CD68	1120	24.17	170.79	1028	30.44	225.23
IDH1	3469	74.86	178.29	2233	66.12	227.41
TMEM119	373	8.05	146.69	151	4.47	138.38
Only CD45	291	6.27	180.07	247	7.31	235.24
No act.	565	12.19	196.53	503	14.89	240.79

Table VIII. Results from sample 1A.

	N	**N [%]**	$\bar{b}_{stain}$	**N**	**N [%]**	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	1824	14.88	108.09	2145	18.92	120.47
CD3	72	0.58	111.83	44	0.38	83.67
CD45	11012	89.84	155.87	9229	81.38	171.82
CD45CD3	72	0.58	111.83	44	0.38	83.67
CD45CD68	8068	65.82	148.09	5888	51.92	174.30
CD45CD68CD163	1687	13.76	108.46	1941	17.11	125.20
CD45TMEM119	8685	70.86	144.57	6659	58.72	144.30
CD68	8406	68.58	149.18	6050	53.35	174.49
IDH1	11866	96.81	160.73	10996	96.97	176.97
TMEM119	9196	75.02	145.51	7458	65.77	139.27
Only CD45	18	0.15	103.35	39	0.34	14.50
No act.	71	0.57	37.53	84	0.74	22.38
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	2121	26.78	137.13	2964	33.21	92.13
CD3	16	0.20	129.07	109	1.22	60.97
CD45	6298	79.52	159.74	8017	89.85	103.51
CD45CD3	16	0.20	129.07	99	1.11	59.02
CD45CD68	4193	52.94	166.53	6155	68.97	95.02
CD45CD68CD163	1743	22.01	134.27	2324	26.04	86.39
CD45TMEM119	3554	44.87	143.43	7630	85.51	96.43
CD68	4731	59.73	166.06	6799	76.19	100.80
IDH1	7242	91.43	173.22	8610	96.49	105.54
TMEM119	3692	79.52	143.94	8327	93.32	98.19
Only CD45	135	1.70	172.80	8	0.09	483.18
No act.	127	1.60	127.78	36	0.4	439.24

Table IX. Results from sample 1B.

	N	N [%]	$\bar{b}_{stain}$	N	N [%]	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	12	0.18	96.65	12	0.21	67.86
CD3	1	0.02	327.94	3	0.05	31.94
CD45	4165	63.06	208.82	3417	59.82	239.71
CD45CD3	1	0.02	327.94	3	0.05	31.94
CD45CD68	855	12.94	193.93	408	7.14	210.76
CD45CD68CD163	10	0.15	90.68	7	0.12	68.76
CD45TMEM119	2288	34.64	199.10	2049	35.87	230.33
CD68	1219	18.46	202.83	677	11.85	220.99
IDH1	3084	46.69	235.02	3620	63.37	249.21
TMEM119	2581	39.07	199.02	2527	44.24	232.32
Only CD45	740	11.20	206.30	457	8.00	242.75
No act.	866	13.11	208.22	639	11.18	253.64
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	10	0.15	97.32	13	0.23	58.46
CD3	23	0.36	190.84	2	0.04	2.62
CD45	4217	65.91	220.21	3361	59.54	177.74
CD45CD3	23	0.36	190.84	2	0.04	2.62
CD45CD68	935	14.61	222.39	416	7.37	180.40
CD45CD68CD163	9	0.14	101.03	7	0.12	89.74
CD45TMEM119	2113	33.03	216.95	1743	30.88	175.08
CD68	1197	18.70	225.41	562	9.96	180.13
IDH1	2849	44.52	184.70	3195	56.60	178.99
TMEM119	2911	45.50	218.75	2490	44.11	176.10
Only CD45	1029	16.08	251.28	626	11.09	183.68
No act.	678	10.59	246.99	612	10.84	188.41

Table X. Results from 2A.

	N	N [%]	$\bar{b}_{stain}$	N	N [%]	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	250	4.46	130.43	168	2.90	109.53
CD3	59	1.05	145.21	64	1.10	153.49
CD45	3908	69.72	186.34	3456	59.77	176.04
CD45CD3	54	0.96	138.47	59	1.02	158.63
CD45CD68	1637	29.20	183.14	1507	26.06	175.19
CD45CD68CD163	122	2.18	117.89	63	1.09	88.83
CD45TMEM119	859	15.33	197.36	515	8.90	173.63
CD68	2132	38.03	184.35	2026	35.04	176.24
IDH1	3633	64.82	186.49	3360	58.11	177.70
TMEM119	1133	20.21	198.90	781	13.50	174.90
Only CD45	679	12.11	198.55	707	12.22	174.99
No act.	431	7.69	195.05	733	12.67	176.67
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	158	5.09	92.99	753	15.11	94.10
CD3	23	0.74	108.32	45	0.90	61.62
CD45	1855	59.8	142.20	3021	60.63	116.55
CD45CD3	13	0.42	100.07	26	0.52	62.32
CD45CD68	677	21.82	137.65	1317	26.43	106.01
CD45CD68CD163	76	2.45	97.16	320	6.42	90.42
CD45TMEM119	359	11.57	135.53	899	18.04	115.77
CD68	1046	33.72	139.03	2044	41.02	108.91
IDH1	913	29.43	158.62	1892	37.97	114.61
TMEM119	578	18.63	142.39	1415	28.40	116.50
Only CD45	669	21.56	141.54	738	14.81	131.70
No act.	556	17.92	150.90	576	11.56	130.31

Table XI. Results from sample 2B.

	N	N [%]	$\bar{b}_{stain}$	N	N [%]	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	44	0.58	40.93	32	0.46	35.43
CD3	5	0.07	151.80	1	0.01	114.03
CD45	3808	50.86	127.09	2890	41.77	157.69
CD45CD3	3	0.04	79.85	0	-	-
CD45CD68	1433	19.14	121.60	1030	14.88	149.44
CD45CD68CD163	26	0.34	30.10	10	0.14	25.50
CD45TMEM119	2553	34.10	126.13	2001	28.92	158.41
CD68	1896	25.32	126.17	1602	23.15	148.56
IDH1	3320	44.34	135.47	1115	16.11	153.40
TMEM119	3183	42.51	129.58	3334	48.18	151.78
Only CD45	480	6.41	126.06	507	7.32	158.77
No act.	1690	22.57	139.55	2002	28.93	139.69
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	58	0.98	54.86	34	0.63	43.40
CD3	5	0.08	120.19	1	0.02	13.00
CD45	2057	34.73	125.03	1368	3945	116.31
CD45CD3	4	0.07	122.77	1	0.02	13.00
CD45CD68	572	9.65	120.96	321	6.04	107.50
CD45CD68CD163	6	0.10	33.90	4	0.08	24.79
CD45TMEM119	1384	23.36	124.52	922	17.35	113.52
CD68	842	14.21	125.58	604	11.36	111.72
IDH1	3143	53.06	134.18	2082	39.18	129.68
TMEM119	2726	46.02	129.70	2312	43.15	120.71
Only CD45	196	3.30	119.89	185	3.48	123.30
No act.	1195	20.17	144.31	1435	27.00	135.17

Table XII. Results from sample 2C.

	N	**N [%]**	$\bar{b}_{stain}$	**N**	**N [%]**	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	1494	38.36	101.29	192	5.27	58.23
CD3	22	0.56	86.75	12	0.33	140.16
CD45	3453	88.67	110.87	3250	89.26	122.93
CD45CD3	22	0.56	86.75	12	0.33	140.16
CD45CD68	2811	72.19	109.78	2764	75.91	121.62
CD45CD68CD163	1288	33.07	100.90	164	4.50	57.27
CD45TMEM119	2563	65.81	111.44	1161	31.88	116.76
CD68	3011	77.32	109.96	2991	82.15	122.10
IDH1	613	15.74	99.27	473	12.99	109.38
TMEM119	2845	73.06	111.83	1292	35.48	116.99
Only CD45	158	4.05	115.38	275	7.55	137.96
No act.	78	2.00	104.74	108	2.97	149.6
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	73	2.15	36.03	1494	38.38	101.29
CD3	10	0.29	135.9	22	0.56	86.76
CD45	2864	84.60	117.36	3453	88.67	110.86
CD45CD3	10	0.29	135.9	22	0.56	86.76
CD45CD68	2407	71.10	116.51	2811	2811	109.78
CD45CD68CD163	65	1.92	36.70	1288	33.08	100.09
CD45TMEM119	652	19.26	107.07	2563	65.82	111.44
CD68	2744	81.06	117.89	3011	77.32	109.95
IDH1	409	12.08	113.43	613	15.74	99.27
TMEM119	740	21.86	107.27	2845	73.06	111.83
Only CD45	315	9.30	124.55	158	4.06	115.38
No act.	130	3.84	130.80	78	2.00	104.74

Table XIII. Results from sample 3A.

	N	**N [%]**	$\bar{b}_{stain}$	**N**	**N [%]**	$\bar{b}_{stain}$
ROI	**1**	**1**	**1**	**2**	**2**	**2**
CD163	2313	23.59	170.48	1562	15.74	228.59
CD3	100	1.02	196.73	105	1.06	300.91
CD45	4215	42.99	181.60	4517	45.52	260.31
CD45CD3	88	0.89	200.31	98	0.98	306.89
CD45CD68	1585	16.16	165.95	1738	17.51	244.29
CD45CD68CD163	715	7.29	152.74	612	6.17	218.53
CD45TMEM119	2062	21.03	178.11	2151	21.67	257.75
CD68	2836	28.93	172.38	2639	26.59	252.53
IDH1	883	9.00	175.69	836	8.42	245.66
TMEM119	4618	47.10	189.25	4702	47.38	274.36
Only CD45	940	9.58	203.04	993	10.00	288.10
No act.	1899	19.36	204.16	2088	21.04	303.93
ROI	**3**	**3**	**3**	**4**	**4**	**4**
CD163	2238	20.07	183.01	1613	14.69	212.36
CD3	267	2.39	191.52	50	0.45	173.48
CD45	4371	39.21	207.73	4069	37.06	220.02
CD45CD3	156	1.39	178.84	17	0.15	141.08
CD45CD68	1154	10.35	154.29	994	9.05	180.05
CD45CD68CD163	484	4.34	124.35	425	3.87	166.64
CD45TMEM119	2204	19.77	212.58	2178	19.83	208.70
CD68	2385	21.39	186.15	2266	20.64	214.13
IDH1	9956	85.72	231.40	9757	88.87	242
TMEM119	5536	47.86	236.19	5181	47.19	238.93
Only CD45	154	1.38	222.42	108	0.98	256.31
No act.	447	4.01	269.98	461	4.19	273.89

Table XIV. Results from 3B.

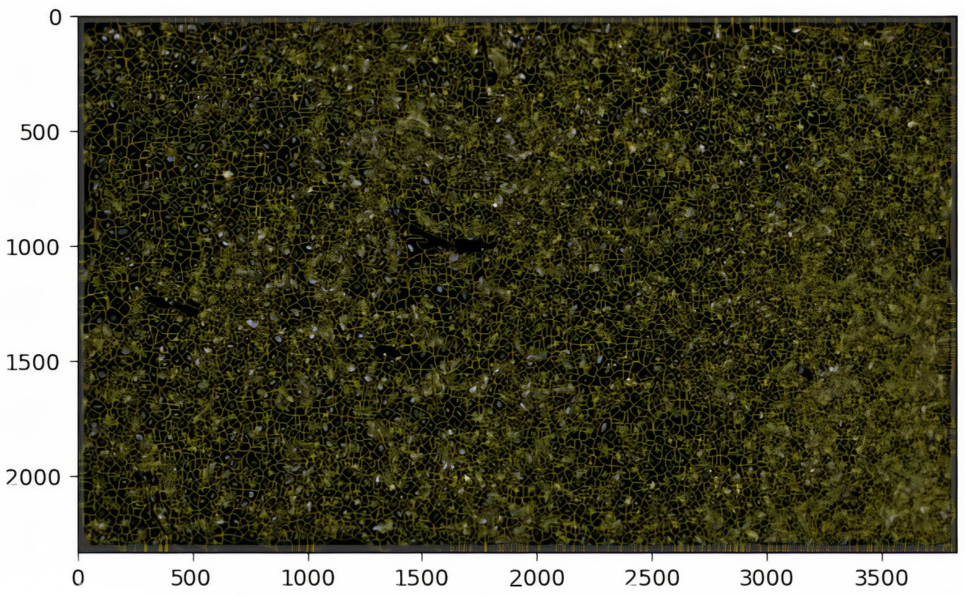

Figure 11. The appearance of cell region segmentation in size of our ROI's. Most of the segmentations are of similar nature, since the ROI's were selected from the middle of tissue segments with dense nuclei populations.

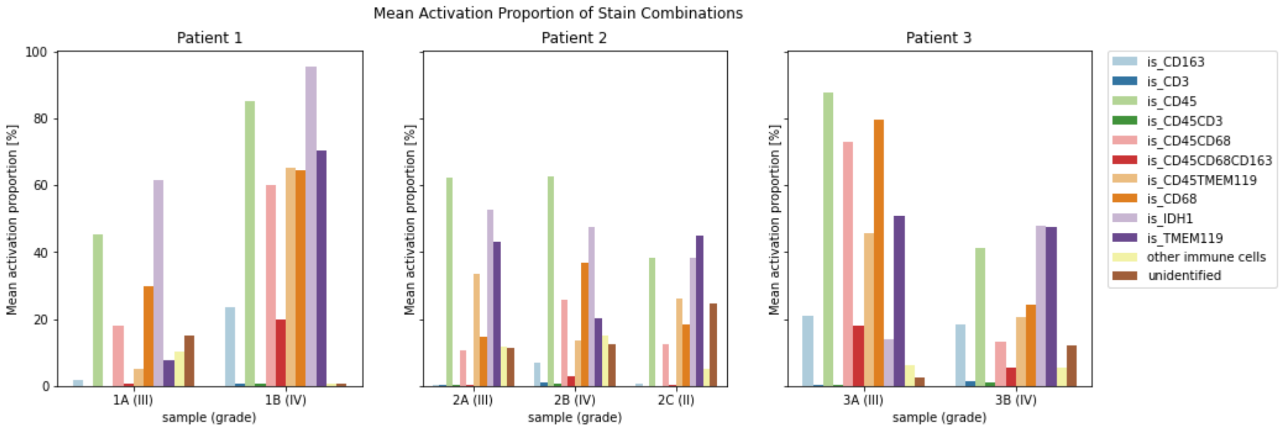

Figure 12. The mean of stain activation proportion from all ROI's per patient per sample.

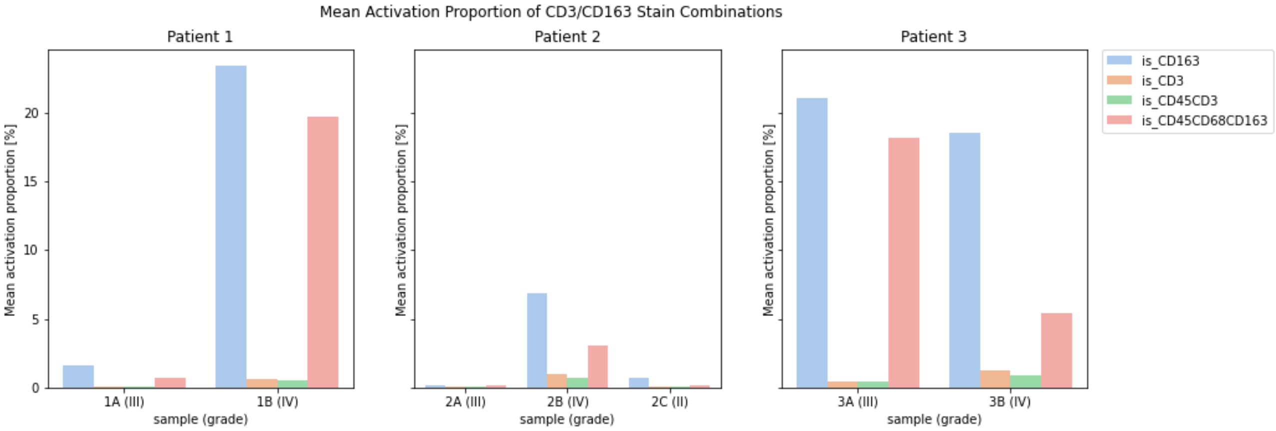

Figure 13. The mean of stain activation proportion from all ROI's per patient per sample for CD3/CD163 combinations.

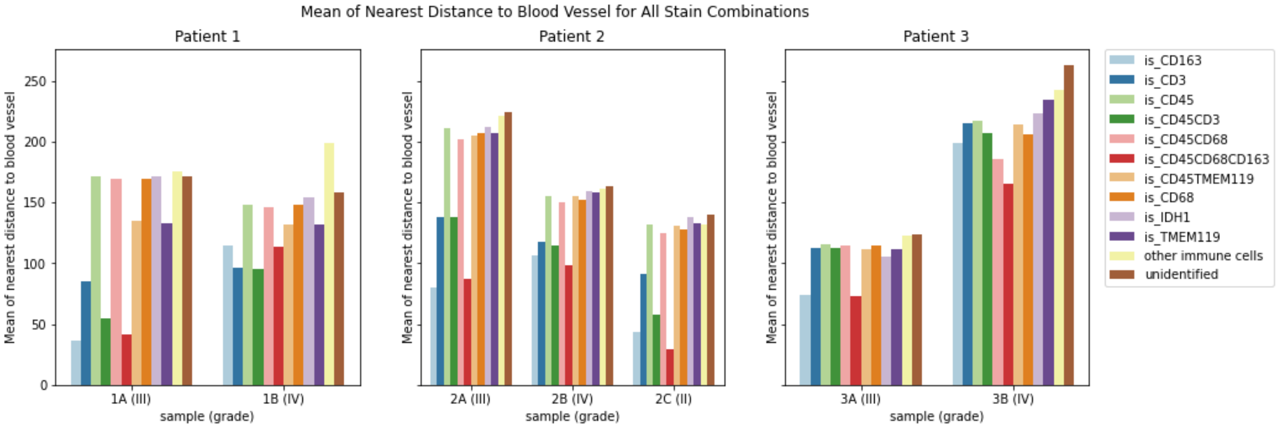

Figure 14. The mean of nearest distance to blood vessel from all ROI's per patient per sample.

4.2 Error analysis

The computational analysis of tumour mIHC tissue samples is only possible through the cooperation of several parties: e.g. a surgeon conducts a biopsy, surgical assistant preserves the tissue, pathologist conducts mIHC/WSI and finally computer vision methods can be applied by an analyst. It is thus apparent, that error margins are high, since each aforementioned step — each with their own constituent processes — contribute to the final outcome.

In terms of the brain tumour biopsy, the sampling process itself is a probable source of error. Only a few tissue segments are extracted from the entire mass and it is thus conceivable that samples are not representative of the true pathological picture. For example, in needle liver biopsies, sampling error has been shown to exist in respect to severity of inflammation, degree of fibrosis and presence of cirrhosis. According to a study by Arie Regev et al., a sampling error may have led to underdiagnosis of cirrhosis in 14.5 % of patients. [47]

Tissue handling can be divided into two main workflows, pre-analytics and analytics. Analytics can be defined as the the processes that are conducted after the sample has been collected and is ready for sectioning for various molecular or histological tests. Conversely, pre-analytics can be considered to include patient identification/tracking, sample collection/transport and embedding into paraffin wax. Whilst analytics has seen substantial development, it is known that pre-analytics steps are largely non-standardized. Poor pre-analytics can lead to many errors in the laboratory, such as mislabelling patient cassettes, improperly fixed tissue, tissue contamination and compromised staining. [48]

As noted in section 3.4 and throughout the thesis, mIHC images have a high level of variability due to several factors. Common culprits include high background noise, over-staining, weak-staining and non-specific binding. These sources of error prove to be a challenging aspect in our analysis: due to variation, it was difficult to find suitable thresholds. In the case that stain channels have high amounts of noise or have been unreliably filtered, the results will not reflect the true profile of the tissues. [35, 42]

We started our methodology with image registration. Three reference points were selected from each sample with care and the registration results were qualitatively successful. It is nonetheless without doubt that registration brings a level of error to our results. Since some channels had very minimal levels of activity, it was hard to compare whether they were successfully registered. Blood vessel outlines were visible in many channels so they acted as our primary reference. Thus, it is possible that there were more severe registration deviations due to differing point selection, that were left undetected. Another noteworthy source of error, is that DAPI registration program output were named very inefficiently, so there was manual work done in assuring that transformation parameters were reapplied correctly, potentially resulting in systematic error.

The selected pre-trained nuclei detection model was indeed produced for dense nuclei areas, which made it fitting to our application. Nuclei detection was qualitatively successful and it set the foundation for all subsequent analysis. Applying watershed segmentation had drawbacks. The size of the cellular area was defined by the distance to other cells and thus algorithm performance at complex morphological sites is poor. Cells located in the borders of inhomogeneous nuclei distributions have exaggerated surface areas, and thus inaccurate activation levels. These cases were filtered out. It is nonetheless likely that our method of segmentation introduced activation errors for our cellular regions.

The program produced data folders for each ROI along with an excel file that contained the analytics regarding activation occurrence. Since the results were manually read from the files and input to this report, there may have been some level of reading error. This could have been alleviated by scripting a solution that automatically produces code for the corresponding LaTeX table. Illustrations 12 - 14 were nonetheless produced programmatically from combined data. Finally, our sampling of 4 ROI's from each case may have resulted in unrepresentative results for each tumour grade. This can be alleviated by increasing the amount of ROI's. The measurements will always have a level of random error.

4.3 Discussion

Our primary finding is that proportion of T cells and M2-macrophage as well as their mean distance to nearest blood vessel are greater in higher tumour grades. This might suggest that T cells and M2-macrophage are of low quantity in near vicinity of blood vessels in low tumour grades, and conversely with higher quantities and more homogeneous distribution in higher tumour grades. It was noted in section 1.3, that research has shown M2-macrophage proportions having positive correlation with tumour grade, which is in line with our results.

Tiihonen mentioned that she had visually made similar observations of the samples, which gives us validation of the results. Whilst visual assessment can give us a general idea, this thesis provides foundation to computational means in automating and quantifying such tissue investigations. Furthermore, our program can be applied to any set of antigens of interest after suitable thresholding — and as the activations are mapped — a plethora of questions can be asked related to cell-cell interactions and tumour grade properties.

Our problem scenario and methodology come with several limitations. As noted in the previous section, there are many sources of error and non-standardized processes related to analysis of tissue. The greatest challenge we faced, unsurprisingly, turned out to be staining variability, as ubiquitously stated in related literature. Our method of using expert knowledge in thresholding gave us sufficient results, yet it is certain that there was a lack of validation in chosen values, since images had different intensity ranges and noise levels.

Our method of segmentation has poor results in heterogeneous tissue areas. In contrast, it is sufficient in densely packed areas, and the ROI's were selected accordingly. Segmented cells are not necessarily realistic representations of cell boundaries, but they are suitable for our means — which is to acquire intensity vectors from given cellular areas. Despite our limitations and low-amount of selected ROI's, the results support pre-existing knowledge.

The issue of staining variability needs to be addressed at both the tissue handling and analytics level; there is a dire need for standardized processes and additional research in generalizability so that models can be applied at a larger scale in a reliable manner. Regarding our methodology, an automated and adaptive solution for image thresolding could be explored. Our program could further be combined with tissue morphology information from H&E staining, for example in tissue detection and automatic tiling.

In terms of our analysis, we had a limited amount of ROI's from each tissue, despite the images being quite large with substantial cell populations. One could further collect data from a larger set of ROI's and conduct proper statistical analysis in supporting hypotheses of micro-environment behaviour regarding tumour progression. In the grand scheme of things, computational utility in mIHC analysis requires a multi-modal system for several tasks, from managing large tabular and image databases to implementing registration modules, automated advanced analysis based on scientific and technical knowledge in tandem as well as visualizing relevant information for medical professionals. If the full pipeline is created in a robust manner, one can imagine commercial-level, semi-automated and advanced mIHC-tissue investigations in clinical use.

References

[1] G.Gopal, C. Suter Crazzolara, L. Toldo et al. (2019), Digital transformation in healthcare - Architectures of present and future information technologies, *Clinical Chemistry and Laboratory Medicine*, **57**(3), 328-335, **URL**: `https://doi.org/10.1515/cclm-2018-0658`

[2] M. Norine Walsh, J. S. Rumsfeld,(2017), Leading the Digital Transformation of Healthcare, *Journal of the American College of Cardiology*, **70**(21), 2719-2722, **URL**: `https://doi.org/10.1016/j.jacc.2017.10.020`

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

[4] M. K. K. Niazi, A. V. Parwani, M. Gurcan, (2021), Digital Pathology and Artificial Intelligence, *The Lancet Oncology*, **20**(5), 253-261, **URL**: `https://doi.org/10.1016/S1470-2045(19)30154-8.`

[5] K. D. Miller, Q. T. Ostrom, C. Kruchko, (2021), Brain and Other Central Nervous System Tumour Statistics, *Cancer Journal for Clinicians*, **71**(5), 381 - 406, **URL**: `https://doi.org/10.3322/caac.21693`

[6] D. F. Quail, J. A. Joyce, (2017), The Microenvironmental Landscape of Brain Tumors, *Cancer Cell*, **31**(3), 326-241, **URL**: `https://doi.org/10.1016/j.ccell.2017.02.009`

[7] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai et. al., (2021), Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy, *Cancer Communications*, **40**(4), 135-153, **URL**: `https://doi.org/10.1002/cac2.12023`

[8] Health Capital Helsinki, (2021), Finnish health startups show great growth potential but funding is urgently needed, Helsinki, Finland, **URL**: `https://healthcapitalhelsinki.fi/health-startup-survey/`

[9] S. Kraus, F. Schiavone, A. Pluzhnikova, (2021), Digital transformation in healthcare: Analyzing the current state-of-research, *Journal of Business Research*, **123**, 557-567, **URL**: `https://doi.org/10.1016/j.jbusres.2020.10.030`

[10] G. Wiederrecht, S. Darwish, A. Callaway, (2020), The healthcare data explosion, *Royal Bank of Canada Capital Markets*, Canada , **URL**: `https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion`

[11] I. Olaronke, O. Oluwaseun, (2016), Big Data in Healthcare: Prospects, Challenges and Resolutions, *2016 Future Technologies Conference (FTC)*, 1152-1157, **URL**: `https://doi.org/10.1109/FTC.2016.7821747`

[12] J. Reponen, T. Vehko, (2021), Terveydenhuollon sähköiset palvelut ja ammattilaistyökalut ovat lisääntyneet, mutta potilastietojärjestelmiä pitää vielä yhtenäistää, *The Finnish Institute for Health and Welfare (THL)*, **URL**: `https://thl.fi/fi/-/terveydenhuollon-sahkoiset-palvelut-ja-ammattilaistyokalut-ovat-lisaantyne`

[13] S. W. Jahn, M. Plass, F. Moinfar, (2020), Digital Pathology: Advantages, Limitations and Emerging Perspectives, *Journal of Clinical Medicine*, **9**(11), 3697, **URL**: `https://doi.org/10.3390/jcm9113697`

[14] S. Heffner, O. Colgan, C. Doolan, Digital Pathology, Illinois, United States **URL**: `https://www.leicabiosystems.com/knowledge-pathway/digital-pathology/`

[15] F. Ghazvani, A. Evans, A. Madabhushi, M. Feldman, (2012) Digital Imaging in Pathology: Whole-Slide Imaging and Beyond, *Annual Review of Pathology: Mechanisms of Disease*, **8**, 331-3359, **URL**: `https://doi.org/10.1146/annurev-pathol-011811-120902`

[16] L. Barisoni, K. J. Lafata, S. M. Hewitt et al., (2020), Digital pathology and computational image analysis in nephropathology, *Nature Reviews Nephrology*, **16**, 669-685, **URL**: `https://doi.org/10.1038/s41581-020-0321-6`

[17] R. Gupta, T. Kurc, A. Sharma et al., (2019) The Emergence of Pathomics, *Current Pathobiology Reports*, **7**, 73-84, **URL**: `https://doi.org/10.1007/s40139-019-00200-x`

[18] C. Srinidhi, O. Ciga, A. Martel, (2021), Deep neural network models for computational histopathology: A survey, *Medical Image Analysis*, **67**, 101813, **URL**: `https://doi.org/10.1016/j.media.2020.101813.`

[19] O. Iizuka, F. Kanavati, K. Kato et al., (2020), Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours, *Scientific Reports*, **10**, 1504, **URL**: `https://doi.org/10.1038/s41598-020-58467-9`

[20] D. Shen, G. Wu, H. Suk, (2017), Deep Learning in Medical Image Analysis, *Annual Review of Biomedical Engineering*, **19**, 221-248, **URL**: `https://doi.org/10.1146/annurev-bioeng-071516-044442`

[21] C. E. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall,(2020), Activation Functions: Comparison of Trends in Practice and Research for Deep Learning, **URL**: `https://doi.org/10.48550/arXiv.1811.03378`

[22] D. Choi, C. J. Shallue, Z. Nado et al., (2020), On Empirical Comparisons of Optimizers for Deep Learning, *ICLR 2020 Conference Blind Submission*, **URL**: `https://doi.org/10.48550/arXiv.1910.05446`

[23] M. Valkonen, G. Högnäs, G. S. Bova, P. Ruusuvuori, (2021), Generalized Fixation Invariant Nuclei Detection Through Domain Adaption Based Deep Learning, *IEEE Journal of Biomedical and Health Informatics*, **5**, 1747-1757, **URL**: `https://doi.org/10.1109/JBHI.2020.3039414`

[24] R. Remark, T. Merghoub, N. Grabe et al., (2016), In-depth tissue profiling using multiplexed immunohistochemical consecutive staining on single slide, *Science Immunology*, **1**(1), 6925, **URL**: `https://doi.org/10.1126/sciimmunol.aaf6925`

[25] P. Hofman, C. Badoual, F. Henderson, (2019), Multiplexed Immunohistochemistry for Molecular and Immune Profiling in Lung Cancer — Just About Ready for Prime-Time?, *Cancers (Basel)*, **11**(3), 283, **URL**: `https://doi.org/10.3390/cancers11030283`

[26] D. Schaffer, S. Willerth, (2017),Scaffold Materials for Human Embryonic Stem Cell Culture and Differentiation, *Comprehensive Biomaterials II*, **5**, 129-153, **URL**: `https://doi.org/10.1016/B978-0-12-803581-8.10111-0`

[27] NCI cancer term dictionary, (2022), *National Cancer Institute*, **URL**: `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/immunohistochemistry`

[28] M. M. Bolognesi, M. Manzoni, C. R. Scalia, (2017), Multiplex Staining by Sequential Immunostaining and Antibody Removal on Routine Tissue Sections, *Journal of Histochemistry & Cytochemistry*, **65**(8), 431-444., **URL**: `https://doi.org/10.1369/0022155417719419`

[29] M. M. Bolognesi, F. M. Bosisio, L. Marcelis et. al., (2019), Multiple Iterative Labeling by Antibody Neodeposition (MILAN), *Protocol Exchange*, **5**, **URL**: `https://doi.org/10.21203/rs.2.1646/v5`

[30] A. Tiihonen, (2020), *Personal notes*, Multiplex IHC staining protocol

[31] F. Hoff, How to Prepare Your Specimen for Immunofluorescence Microscopy, *Leica Microsystems*, Illinois, United States, **URL**: `https://www.leica-microsystems.com/science-lab/how-to-prepare-your-specimen-for-immunofluorescence-microscopy/`

[32] Johns Hopkins Medicine, (2022), Baltimore, USA, Brain Tumour in Children, **URL**: `https://www.hopkinsmedicine.org/health/conditions-and-diseases/brain-tumor/pediatric-brain-tumors`

[33] K. Aldape, K. M. Brindle, L. Chesler, (2019), Challenges to curing primary brain tumours, *Nature Reviews Clinical Oncology*, **8**, 509-520, **URL**: `https://doi.org/10.1038/s41571-019-0177-5`

[34] M. D. Sorensen, R. H. Dahlrot, H. B. Boldt et. al., (2018) Tumour-associated microglia / macrophages predict poor prognosis in high-grade gliomas and correlate with an aggressive tumour subtype, *Neuropathology and Applied Neurobiology*, **2**, 185-206, **URL**: `https://doi.org/10.1111/nan.12428`

[35] D. J. Fassler, S. Abousamara, R. Gupta et. al., (2020), Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images, *Diagnostic Pathology*, **15**, 100, **URL**: `https://doi.org/10.1186/s13000-020-01003-0`

[36] Y. B. Hagos, P. L. Narayanan, A. U. Akarca et. al., (2019), ConCORDe-Net: Cell Count Regularized Convolutional Neural Network for Cell Detection in Multiplex Immunohistochemistry Images, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)*, **11764**, **URL**: `https://doi.org/10.48550/arXiv.1908.00907`

[37] Y. Wang, Y. G. Wang, C. Hu et. al., (2021), Cell graph neural networks enable the digital staging of tumor microenvironment and precise prediction of patient survival in gastric cancer, *medRxiv*, **URL**: `https://doi.org/10.1101/2021.09.01.21262086`

[38] Cytomine, Open-source rich internet application for collaborative analysis of multi-gigapixel images, Liege, Belgium, **URL**: `https://cytomine.be/`

[39] SchedMD,(2022), Slurm Workload Manager Overview, Utah, United States, **URL**: `https://slurm.schedmd.com/overview.html`

[40] Tampere University ITC Wiki, (2022), TUNI Narvi Cluster, Tampere, Finland, **URL**: `https://tuni-itc.github.io/wiki/Technical-Notes/tuni-narvi-cluster/`

[41] Leow Wee Khang, (2022), CS4243 Computer Vision and Pattern Recognition, National University of Singapore, **URL**: `https://www.comp.nus.edu.sg/~cs4243/lecture/register.pdf`

[42] L. Pozzi,(2022), Artifacts in IHC, The usual suspects - part I, Eching, Germany, BIOZOL, **URL**: `https://www.biozol.de/en/techniques/Artifacts-in-IHC`

[43] U. Schmidt, M. Weigert, C. Broaddus et. al., (2018), Cell Detection with Star-convex Polygons, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **11071**, **URL**: `https://doi.org/10.1007/978-3-030-00934-2_30`

[44] O. Ronneberger, P. Fischer, T. Brox, (2015), U-Net, Convolutional Networks for Biomedical Image Segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **9351**, **URL**: `https://doi.org/10.1007/978-3-319-24574-4_28`

[45] A. Bieniek, A. Moga, (1999), An efficient watershed algorithm based on connected components, *Pattern Recognition*, **33**(6), **URL**: `https://doi.org/10.1016/S0031-3203(99)00154-5`

[46] A. S. Kornilov, I. V. Safonov, (2018), An Overview of Watershed Algorithm Implementations in Open Source Libraries, *Journal of Imaging*, **4**(10), **URL**: `https://doi.org/10.3390/jimaging4100123`

[47] A. Regev, M. Berho, L. J. Jeffers, (2002), Sampling Error and Intraobserver Variation in Liver Biopsy in Patients With Chronic HCV Infection, *The American Journal of Gastroenterology*, **10**, 2614-8, **URL**: `https://doi.org/10.1111/j.1572-0241.2002.06038.x`

[48] D. R. Bauer, M. Otter, D. R. Chaffin, (2018), A New Paradigm for Tissue Diagnostics: Tools and Techniques to Standardize Tissue Collection, Transport, and Fixation, *Current Pathobiology Reports*, **6**(2), 135-143, **URL**: `https://doi.org/10.1007/s40139-018-0170-1`