Deep learning for histology-based cancer research and diagnostics

Thagaard, Jeppe

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Thagaard, J. (2021). *Deep learning for histology-based cancer research and diagnostics*. Technical University of Denmark.

Deep learning for histology-based cancer research and diagnostics

Jeppe Thagaard

Kongens Lyngby 2021

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

Nearly 20 million people around the world were diagnosed with cancer in 2020 causing almost 10 million lives to be lost. Female breast cancer has surpassed lung cancer as the most frequent type with 2.3 million women diagnosed with breast cancer every year. The total global cancer burden is expected to rise by 47% from 2020 to be 28.4 million cases in 2040. The machine learning community has an obligation to use the power of deep neural networks to look for novel and sustainable solutions that make a true impact on the field of pathology - the rock bed of cancer research and diagnostics. Currently, the leaps in development are often confined to the proof-of-concept stage, never reaching the end-users. The main challenges are related to both complex diagnostic regimes, lack of standardization, and the cost of obtaining training data. These aspects make it difficult to build algorithms that generalize into the clinical domain.

The goal of this thesis is to investigate some of the challenges of bringing algorithms into real-world settings in pathology. By studying realistic shifts in data distributions, we show that deep neural networks can generalize to and provide reliable uncertainty estimates within the cancer indication it was trained on. On the contrary, they fail to report rare abnormalities, and other systems need to inspect incoming data for signs of significant changes in input distributions. Moreover, we demonstrate that it is possible to create an automatic computational approach to quantify the tumor infiltrating lymphocytes (TILs) that could help in standardizing the prognostic assessment. In turn, this can support clinicians in treatment decisions to provide better patient outcomes. Meanwhile, we also investigate a more objective and scalable method to create training labels. Finally, we assess some of the remaining challenges of using machine learning, and how these can be overcome with further research.

Summary (Danish)

Omtrent 20 millioner mennesker blev i 2020 diagnosticeret med cancer på verdensplan, hvilket resulterede i tab af næsten 10 millioner menneskeliv. Brystcancer blandt kvinder har overgået lungecancer som den hyppigste cancerform med 2,3 mio. diagnosticerede kvinder årligt. Den globale cancerbyrde forventes at stige med 47% fra år 2020 til 28,4 millioner cases i 2040. Machine learning feltet er forpligtet til at anvende mulighederne ved dybe neurale netværk til at afsøge nye og vedvarende løsninger, som har en reel indvirkning på patologifeltet – fundamentet for cancerforskning og -diagnostik. For nuværende, er udviklingen ofte begrænset til prototyper, som aldrig når slutbrugeren. De største udfordringer er relaterede til komplekse diagnostiske regimer, manglende standardisering samt omkostninger forbundet med indsamling af træningsdata. Disse aspekter vanskeliggør udviklingen af algoritmer, der kan generalisere tilstrækkelig med henblik på klinisk anvendelse.

Formålet med denne afhandling er at undersøge nogle af de udfordringer, der ligger i at overføre algoritmer til virkelighedens patologi-verden. Ved at undersøge realistiske forskydninger i datafordelingen, viser vi, at dybe neurale netværk kan generalisere og bidrage med pålidelige usikkerhedsestimater inden for den cancerform, de er trænet på. Derimod er de ikke i stand til at rapportere sjældne abnormiteter, hvorfor andre systemer må inspicere tegn på signifikante ændringer på fordelingen af indkomne data. Vi demonstrerer herudover, at det er muligt at skabe en fuldt automatisk computerbaseret tilgang til kvantificering af tumorinfiltrerende lymfocytter, som kan bidrage til en standardisering af prognostiske vurderinger. Dette kan støtte klinikerne til at identificere den mest hensigtsmæssige behandlingsstrategi for den enkelte patient. På samme tid undersøger vi en mere objektiv og skalerbar metode til at skabe træningseksempler. Slutteligt, afsøger vi nogle af de resterende udfordringer i anvendelsen af machine learning, og hvordan disse kan overkommes ved hjælp af yderligere forskning.

Preface

This thesis was prepared at the Section for Visual Computing under the Department of Applied Mathematics and Computer Science, at the Technical University of Denmark (DTU) and Visiopharm A/S in fulfilment of the requirements for acquiring a doctor of philosophy (Ph.D.) degree at DTU with an emphasis on computer vision and machine learning.

Professor Anders B. Dahl and Professor Søren Hauberg supervised the project at DTU while Thomas Ebstrup acted as a company supervisor at Visiopharm A/S. The project was funded by Innovation Fund Denmark (grant number 8053-00008B) and Visiopharm A/S.

The thesis consists of three papers (2 published and 1 to be submitted), which are each presented with a thorough introduction. All papers are appended in this thesis.

Lyngby, 31-August-2021

Jeppe Thagaard

Acknowledgements

First and foremost I would like to thank my three supervisors, Professor Anders B. Dahl, Professor Søren Hauberg and Thomas Ebstrup, for providing highly professional scientific guidance and collaboration throughout the research. This has led to a PhD study that has been both a fun and challenging ride through a pandemic.

I would also like to thank my research collaborators: Elisabeth Specht Stovgaard, Eva Balslev, Bianca Grosser and Ralf Huss and my Visiopharm colleagues: Johan Doré, Line Vognsen, Fabian Schneider, Rasmus Lyngby, Henrik Høeg and Michael Grunkin (and a lot more Visiopharm family) for interesting discussions and insights throughout the research.

Thanks to the Visiopharm A/S and Innovation Fund Denmark for funding the research.

Last but not least, I would like to thank Cecilia Søbye Petersen, for her eternal support throughout these studies. As always, none of this was possible without you by my side.

Contributions

Papers included in thesis

A. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J. D., & Dahl, A. B. (2020). Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *Proceedings of 23rd International Conference on Medical Image Computing and Computer Assisted Intervention* (pp. 824-833)

B. Thagaard, J., Stovgaard, E.S., Vognsen, L.G., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Vincentz, R.E., Jepsen, R.K., Roslind, A., Kümler, I., Nielsen, D., & Balslev, E. (2021) Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers. In *Cancers* 13(12):3050.

C. Thagaard, J., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Roslind, A., Nielsen, D., Balslev, E., Salgado, R., ..., & Stovgaard, E.S. (2021) Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer. To be submitted.

Papers not included in thesis

i. Amgad, M., Stovgaard, E. S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., ... & Cooper, L. A. (2020). Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group. In *NPJ breast cancer*, 6(1), 1-13.

ii. Vossen, D., Thagaard, J., Schneider, F., Sørensen, R.N. , Doré, J., Abels, E., Lowe, A., Vyberg, M., Grunkin, M. (2022). AI-Driven Precision Pathology: Challenges and Innovations in Tissue Biomarker Analysis for

Diagnosis. To appear in Artificial Intelligence Applications in Human Pathology. *WORLD SCIENTIFIC (EUROPE)* . DOI:10.1142q0336

iii. Huss, R., Schmid, C., Manesse, M., Thagaard, J., & Maerkl, B. (2021). Immunological tumor heterogeneity and diagnostic profiling for advanced and immune therapies. In *Advances in Cell and Gene Therapy.*

Abbreviations

ASCO	American Society of Clinical Oncology
ASPP	Atrous Spatial Pyramid Pooling
CAP	College of American Pathologists
CNN	Convolutional Neural Network
CTA	Computational Tumor Infiltrating Lymphocytes Assessment
DCIS	Ductal Carcinoma In Situ
ECE	Expected Calibration Error
EQA	External Quality Assessment
ER	Estrogen Receptor
ESMO	European Society for Medical Oncology
GAP	Global Average Pooling
H&E	Hematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
HR	Hazard Ratios
HR+	Hormone-positive
IHC	Immunohistochemical
iTILs	Intratumoral Tumor Infiltrating Lymphocytes

LCIS Lobular Carcinoma In Situ

LIS Laboratory Information Systems

mIF Multiplexed Immunofluorescence

OOD Out-Of-Distribution

OS Overall Survival

PDS Physical Double Staining

PH Proportional Hazards

PR Progesterone Receptor

RFS Relapse-Free Survival

SCC Squamous Cell Carcinoma

sTILs Stromal Tumor Infiltrating Lymphocytes

TIL Tumor Infiltrating Lymphocytes

TIL-WG International Immuno-Oncology Biomarker Working Group on Breast Cancer

TME Tumor Microenvironment

TNBC Triple-Negative Breast Cancers

UQ Uncertainty Quantification

VTA Visual Tumor Infiltrating Lymphocytes Assessment

WSIs Whole Slide Images

Contents

Chapter 1

Introduction

Pathology is the cornerstone of biomedical research and clinical cancer diagnostics. With any suspicion of abnormal changes in human tissues, e.g. a feeling of a lump in a breast or a nodule on a lung x-ray, a tissue biopsy is always acquired to examine the cells and tissue structures microscopically before making a final diagnosis. Such diagnostic discipline relies heavily on the expert training of specialized doctors (pathologists) to recognize patterns, and interpret patterns in the wider context of each patient. Even though it is a necessary prerequisite for any reliable treatment of diseases such as cancer and is, by far, the least expensive diagnostic procedure, reproducibility among pathologists is not optimal. Meanwhile, there is an increasing shortage of pathologists, their workload is worsening as a result of larger numbers of cases, and the requirements are increasing for more extensive diagnoses to identify the optimal treatment for patients in the age of precision medicine.

Machine learning-based algorithms have early on been destined (and promised) to change the medical practice of pathology. With the last decade's recent progress in the digitization of data acquisition, image-based learning algorithms, and computing infrastructure, the realization of this impact is gradually happening - also in areas that were previously thought to be exclusive to medically trained human experts. However, many applications still never leave the academic prototype stage or proof-of-concept product stage, and therefore do not deliver value to the pathologists, the patients, or the healthcare system as a whole. Whereas many technical challenges have been overcome in a grand challenge setting with a fixed training and testing set, several real-world hurdles still have to be overcome on the road towards clinical usefulness.

The generalizability of algorithms is still of the biggest challenges in pathology, mainly due to the lack of standardization causing a high degree of variability present in data. This makes it inherently difficult to create large enough datasets that encapsulate all important characteristics needed to make them representative of the data encountered in clinical practice. There have been many proposed solutions to solve this ranging from removing variability at test time with normalization to introducing more variability at training time with data augmentation [1]. These methods definitely help to increase generalization but do not take into account the high variability of tissue, and especially the endless possible phenotypical morphology of cancer cells. Therefore, they do not replace the benefits of incorporating the natural data variability of large multi-institutional datasets.

Assume now, no matter the impracticability, that a high enough number of laboratories are included in a dataset that should be representative of real-world settings. First, how do we use the data in such a way that we create a computational algorithm that brings clinical utility? In the dawning era of personalized medicine, advanced biomarkers have become critical gatekeepers to the use of new personalized prognostics and treatment options. Naturally, this can result in complex and ambiguous scoring guidelines that restrict future biomarker candidates to progress into the routine clinical management of cancer patients. Therefore, addressing the increased complexity and ambiguousness in clinical assessments of predictive biomarkers is a critical component to ensure reliable reporting. Moreover, how do we obtain the training data needed to deliver such an algorithm? Especially in pathology, where pathologists have limited time, and when do they do, they suffer from reproducibility issues on certain labeling tasks. Collecting training labels also remains one of the main barriers of scaling of algorithms, where especially label consistency (removing noisy data) is the most important aspect to get right in an increasing trend of data-centric machine learning development. Imagine next, that we have found a way to solve a valuable task and labeled all the data in such a way that we can develop and deploy the algorithm. How do we know that we have included all the variability representative of the type of data that is encountered in clinical practice? This brings us to the final aspect; can we possibly know that? No matter how much data, we include into the development dataset, there will always be situations where there are inputs not seen before or there is not necessarily a definitive answer. When a pathologist in unsure on a cases, they can ask of a second opinion or further investigate the case with more advanced tests. However, in these situations, we want to know how a deployed model behaves; does it fail silently and provide false information to the pathologists or can it let us know when there is something that it doesn't know? These aspects lead to the aim and objectives of this thesis.

1.1 Aim & Objectives

The primary aim of this Ph.D. project is to investigate some of the unknowns and challenges of bringing algorithms into real-world settings in pathology as mentioned above. The focus has therefore been on 3 different aspects listed below:

- Objective 1: Investigate what happens under deployment settings where unknown changes to the data distribution might occur unknowingly, and do the current deep neural network models know what they do not know? (related to the paper A)
- Objective 2: Investigate what it takes to translate an emerging biomarker into a fully automated computational-derived biomarker? (related to paper B and C)
- Objective 3: Investigate ways to create better training data with less effort for pathologists (related to paper B)

1.2 Structure of thesis

We expect the reader to read the entire thesis including our research papers appended in **Appendix A, B and C** which we consider our main work of this thesis. Hence, the chapters are meant to set up these contributions and are not a reformatting of our papers. The thesis is organized as follows:

Chapter 2 sets the scene of the thesis by providing background knowledge on breast cancer pathology and diagnosis as its the main type of cancer-related to our research, as well as introducing computational pathology and deep neural networks.

Chapter 3 introduces the underlying methods of uncertainty quantification, and how this is influenced by certain aspects of realistic changes in data that can occur in a day-to-day pathology laboratory. It lays the foundation of our work on uncertainty quantification and dataset shifts (c.f. **Appendix A**).

Chapter 4 introduces a pathology guideline for quantifying how immune cells infiltrate a certain subtype of breast cancer and presents the technical methods used to develop and evaluate a fully automated algorithm that adheres to such guidelines. In addition, the chapter summarizes the key contributions and findings of **Appendix B** and **Appendix C**.

Chapter 5 summarizes and discusses the implications of the work in detail. In addition, the chapter outlines the opportunities for future research in the area.

Chapter 6 draws the final conclusions of the thesis.

Chapter 2

Background

In this chapter, we provide general background knowledge of breast cancer pathology and computational pathology to set the Ph.D. project into context. We also introduce deep neural networks to familiarize the reader with notations that are used in later chapters.

2.1 Breast cancer pathology

Breast cancer is the most prevalent cancer worldwide with more than 2.3 million women receiving a diagnosis and 685.000 deaths in 2020 [2] and is the leading cause of death from any type of cancer among women age 20-39 [3]. It originates in the epithelial cells of the ducts (85%) or lobules (15%) in the glands of the breast [4]. Its progression starts from normal epithelial cells evolving into malignant invasive carcinoma via a pre-invasive in-situ stage. In-situ refers to lesions, where the cancerous cells are kept within the ducts (ductal carcinoma in situ (DCIS)) or lobules (lobular carcinoma in situ (LCIS)) [5], see Figure 2.1. While the clinical and pathological importance of DCIS/LCIS lesion is still an open research question, the current practice defines a lesion as malignant when it progresses and invades the surrounding breast tissue. The survival chances are generally high as the treatment options can be highly effective, mainly due to advances since the 1980s in the earlier diagnosis through screening programs and better profiling of breast cancer subtypes that allows for more targeting treatments [4].

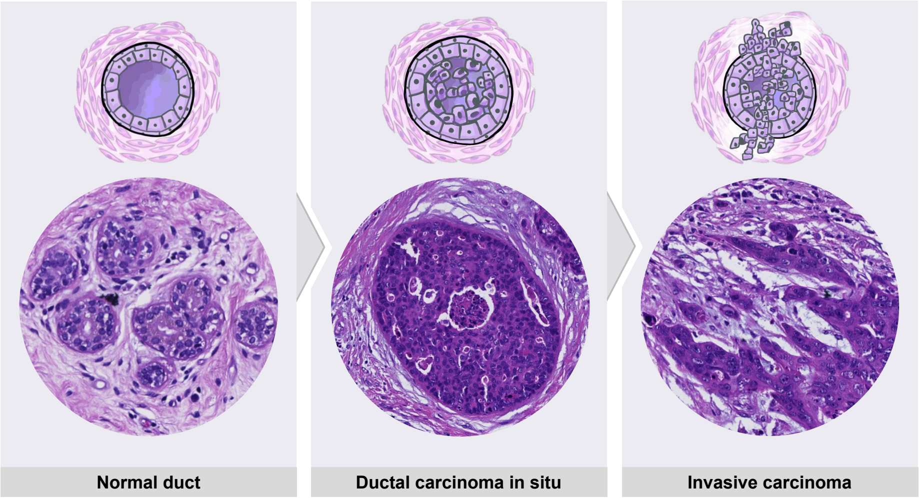


Figure 2.1: Progression of breast cancer stages from normal epithelium to invasive carcinoma where ductal carcinoma in situ and invasive carcinoma are considered abnormal ducts. Here stained with hematoxylin and eosin (H&E). Top pane illustration graphics from [5]

2.1.1 Current diagnostic paradigm

Currently, the diagnosis is based on thorough examinations of cells or tissue sampled by a core biopsy or fine-needle aspiration. Simplified, there are three high-level diagnostic aspects relevant for this thesis; the determination of (i) the pathologic stage, (ii) histologic grade, and (iii) intrinsic subtype of cancer. The goal is to obtain data that help a multi-disciplinary team to give the patient a prognosis and provide the best treatment options possible.

Pathologic stage uses information about the primary tumor (T), involvements of lymph node metastases (N), and distant metastases (M) to provide a score of the stage. The pathological T-score (pT) is based on the size of the tumor, and the localization of the cancer cells (skin or chest wall). The pN-stage is determined by the number and localization of lymph nodes classified as positive for metastasized cancer cells and the size of metastases. The pM-stage is a binary score based on the confirmed presence of metastases in a distant localization. The TNM system is an internationally used scoring guideline where a higher score means more advanced disease and worse prognosis [5].

Histologic stage is the pathologist's assessment of tubule/gland formation, nuclear pleomorphism, and mitotic count. These features capture how similar the cancer cells organize, look, and proliferate compared to normal cells. Higher grade tumors tend to grow and metastasize more aggressively [5].

Intrinsic subtypes is a more novel classification of breast cancer than only histological appearances and other clinical parameters [6]. Originally, gene expression was used to categorize breast cancers as either luminal A, luminal B, HER2-enriched, or basal-like. The subtyping facilitates both prognostic and predictive information for treatment response, i.e. does a patient respond to a certain treatment, e.g., Herceptin [7]. Today, cheaper and more accessible immunohistochemical (IHC) stains[1] are used as surrogate biomarkers for subtyping; two hormone receptors, estrogen (ER) and progesterone (PR), human epidermal growth factor receptor 2 (HER2), and a proliferation marker (Ki67). By assessing the biomarkers, breast cancer expressing ER or PR are hormone-positive (HR+) while HER2 expressing tumors are simply called HER2+. If cancer does not express any of these biomarkers, they are classified as triple-negative breast cancers (TNBC). Ki67 is an alternative to mitotic counting in histologic grading. See Figure 2.2 for general classification scheme. Not only are these biomarkers associated with prognosis, but there are special treatment options for ER+, PR+, and HER2+ cancers [5]. Similar Ki67 is used clinically to decide which HR+ patients to give chemotherapy [8]. However, around 15% of all breast cancers are TNBC, for whom there, until recently [9], have not been any special medications. In general, luminal A is the most frequent subtype (70%) and also has best prognosis [10]. TNBC (15%) and HER2 enriched (10%) cancers are considered to be more aggressive with a poorer prognosis than the other types of breast cancers [10].

There are lots of details and methods left out in this introduction such as the field of molecular pathology and genetic analysis of cancers, which is not directly relevant for this thesis. The N-stage and histologic grading are relevant for Chapter 3, while Chapter 4 looks closer at the prognosis of TNBC. But for both pathologic and histologic staging and the semi-quantitative assessment of biomarkers, extensive efforts have been made towards standardization. However, manual assessment can be subjective and suffer from inter- and intra-observer variability. For example, pathology review by experts changed the nodal status in 24% of patients [11]. Besides being subject to these factors, pathologists are also experiencing an increased workload due to larger numbers of cases, that also require more extensive diagnoses using highly complex biomarkers to identify the best treatment options for patients.

[1] Sometimes referred to as IHC4

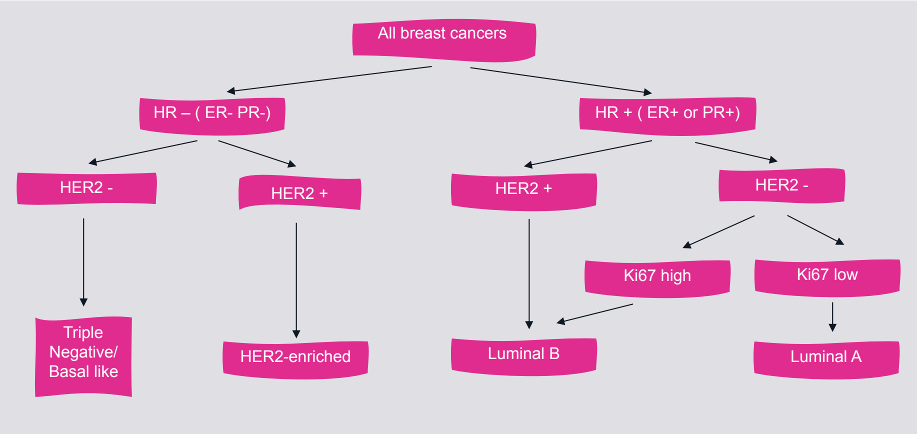


Figure 2.2: Intrinsic subtypes of breast cancer and their IHC classification. Luminal A is the most common subtype, followed by TNBC and Luminal B with HER2-enriched being the least common subtype. Note that the definition of Ki67 might be different depending on regional guidelines.

2.1.2 The immune tumor microenvironment

There is an ongoing shift from mainly investigating the tumor as described above towards focusing on the host in which it exists, often referred to as the immune tumor microenvironment (TME). Specifically, immunotherapy, a certain line of new treatments, has changed the understanding of the uniqueness of each patient's immune TME. Especially, ways a patient's immune system can be unleashed as a novel method to treat cancer. One of the breakthroughs is the understanding of how some immune cells contribute to the anti-tumor response while others promote cancer growth [12]. To a certain extent, this all happens from the bidirectional influence that immune and cancer cells have on each other [13]. Therefore, there is evidence that the characterization of the density, location, and organization of immune cells, the so-called immune landscape, can be used as a surrogate to evaluate a patient's immune response and tumor immunogenicity [14]. The association between one immune cell type, lymphocytes, and survival of patients was noticed almost 100-years ago [15], but then forgotten until the early 1990s [16] where the association between tumor infiltrating lymphocytes (TIL) and outcome in breast cancer was reported. Since then, many studies have studied the prognostic and predictive value of TIL in different breast cancer subtypes, but it was only recently that the clinical validity for early-stage TNBC became well-established through clinical trials [17, 18, 19]. The evaluation of TILs was recommended in the 2019 St. Gallen International Breast Cancer Conference for routine diagnostics of TNBC [20], and in Denmark, the evaluation of TILs is now incorporated in national guidelines as an optional item for TNBC diagnostics. This biomarker is the topic of Chapter 4.

We have left out a lot of recent research and details on the topic of the immune TME, but refer the reader to an overview of TILs from the International Immuno-Oncology Biomarker Working Group on Breast Cancer [14], and Huss et al. (2021) [13] for a review of the relevance to understand the tumor heterogeneity and diagnostic profiling for new therapies. However, there is a hope that recent breakthroughs in computerized image analysis could potentially address these aspects of the immune TME.

2.2 Computational precision pathology

In this section, we briefly introduce key concepts and terminology when using image analysis to analyze tissue sections - often referred to as computational pathology. First, we give an overview of the data generation process before introducing the different tasks that computational pathology can be applied to. Here, we also provide a perspective on the requirements to standardization and automation are needed to deliver effective results for specific diagnostic tasks.

2.2.1 Understanding the data

The procedure for obtaining the tissue sections for microscopic investigates is complex with many individual manual and automated steps. First, tissue sections are prepared by the following steps: (i) sampling the tissue removed from the body for diagnosis. (ii) fixating it to preserve the state of the tissue. (iii) embedding the tissue in paraffin. (iv) sectioning the tissue into thin (3-5 µm) sections, and finally (v) mounted onto glass slides. Until here, each step involves many variables that, if changed or are different between two patients, might introduce variability in the final data. We will refer to these as *preanalytical variables*. After step (v), the tissue is colorless without much information and the slide needs to be stained and then digitized. We will refer to these two factors as *analytical variables* that also suffer from the lack of standardization and might also introduce variability in the data.

Staining introduce contrast and highlight important features of the otherwise transparent tissue such as special stains, multiplexed immunofluorescence (mIF), IHC, and hematoxylin, and eosin (H&E). We will only cover the two later here. H&E is the most common staining in pathology used for at least a century to create contrast between various tissue components [21]. Hematoxylin stains cell nuclei blue/purple, and eosin stains non-nuclear components in different shades of pink. H&E staining is non-specific, meaning that it stains most of the cells similarly and does not target a specific protein. This is the basic principle behind IHC that utilizes antibody-antigen specificity to selectively stain specific chemical compounds or molecules e.g., receptor proteins. Most commonly, an antibody is tagged with an enzyme that catalyzes specific coloring [22], making it possible to capture the signal of specific cellular components within a cell or tissue. H&E is considered the golden standard for many diagnostic tasks, while IHC is used for certain biomarkers (e.g., IHC4) or to investigate the origin of cancer cells.

Whole slide imaging is the process of creating digitized versions of glass slides called whole slide images (WSIs), giga-pixel images stored in a pyramidal multi-resolution file structure, see Figure 2.3. Even though the technology has been around for 30 years, it is only recently that a minority of hospitals are starting to digitize their glass after recent advances in scanning speed, quality, and cost. Also, as part of the data generation process, there is a lack of standardization. There are efforts towards DICOM [23], but most different manufacturers still have proprietary WSI formats, where acquisition parameters such as size, contrast enhancement and gamma adjustment are format specific. Even with these challenges, there are many advantages of digitization such as remote diagnostics during a pandemic. However, it is a requirement before computational pathology can be applied to assist the pathologists on certain tasks.

2.2.2 Standardization and automation

There are 4 main use cases where image analysis can aid pathologists and pathology laboratories:

1. Optimize routine workflow in pathology: Tedious and time-consuming tasks that require a high level of accuracy such as detection of metastases in lymph node sections. Here, cases could be triaged by an automatic detection system so the pathologist can focus on the most important tasks, and sign-off cases faster.

2. Predict outcome and treatment response: Standardized quantification of biomarkers that provide value for patient treatment management.

3. Enable scientific insights: Infer new biological insight from images not directly known to pathologists such as genetic alterations and spatial tumor

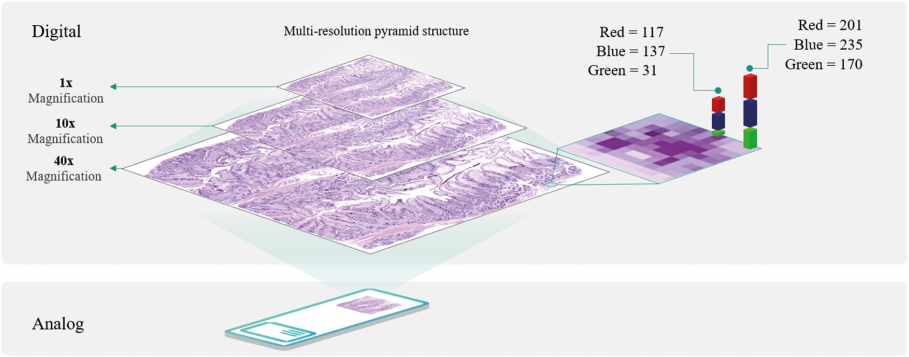


Figure 2.3: A typical WSI used for diagnostic reading is scanned at 40x (0.25 $\mu m/pixel$) or 20x (0.50 µm/pixel) magnification, generating a giga-pixel image (200.000x100.000 pixels) stored in a multi-resolution pyramid structure ensuring image access e.g., zooming, panning, etc. For comparison, a single WSI includes the same amount of data as more than 1500 modern smartphones images.

heterogeneity.

4. Improve the quality of preanalytical and analytical variables: Detect, quantify, and provide feedback of quality issues such as stain proficiency [24] or artifacts [25].

There are different value-propositions but also requirements for these use-cases, also from a regulatory perspective [26]. This thesis mainly focuses on the applications from the first two but also discusses aspects of the other when relevant. One general consideration is the need to create robust and high throughput algorithms that increase standardization of pathology and integrates seamlessly in the currently established workflow without introducing more manual work. At a high level, there have been three different ways of integrating computational pathology in the workflow:

- Workflow 1: The pathologist should manually draw the regions that should be analyzed, and then wait for the analysis to complete. Both these aspects could be time-consuming depending on the complexity of the analysis. The final accuracy depends on the precision of the manual drawing, hence limits standardization.

- Workflow 2: Here, the entire slide is pre-analyzed by the algorithm before the pathologist manually draws the regions, hence decreasing the wait time of the analysis. However, the manual drawing is still time-consuming and affects the accuracy similar to workflow 1.

- Workflow 3: The entire slide is pre-analyzed, where the algorithm both outlines the regions, replacing the manual drawing step, and analyzes within those regions. Therefore, the manual time-consuming tasks are completely removed and there is an increase in standardization as all cases are analyzed in the same manner. The purpose of these types of algorithms is that they provide the full analysis results for when the pathologist opens the case for the first review.

There are differences depending on the use-case, but the general trend is towards workflow 3 to ensure a higher degree of automation and standardization. The underlying advancement in technology that enables this trend is deep neural networks with their ability to learn patterns directly from images. In the next section, we will introduce the topic of deep neural networks.

2.3 Deep neural networks

In this section, we give a brief introduction to deep neural networks with a special focus on the notation and variants used in our research. This section is not an in-depth walk-through of all aspects related to this field. We only introduce models based on gradient descent optimization algorithms that in a supervised setting use a set of N training examples of inputs $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ and labels $\mathbf{y} = \{y_1, y_2, ..., y_N\}$. We denote the model's learnable parameters θ. The term *deep* in deep neural network refers to structural stacking of layer functions, where each layer is a linear model $h(\cdot)$ with its own learnable parameters θ_l combined with non-linear activation function $\sigma(\cdot)$. Let $\mathbf{z}$ be the final-layer output, then a deep neural network with L layers can be written as:

$$\mathbf{z} = \mathbf{z}_\theta(\mathbf{x}) = h_l(h_{l-1}(...h_2(h_1(\mathbf{x}))...) \tag{2.1}$$

For most classification tasks, we pass the final-layer output a softmax activation function $\sigma_y(\cdot)$ to give the model's predictions $p(y|x, \theta)) = \sigma_y(\mathbf{z}) \in [0, 1]^K$,

$$\sigma_y(\mathbf{z})_i = \frac{e_i^z}{\sum_{j=1}^K e_j^z} \tag{2.2}$$

for K output classes. That is, the deep neural network is a function that maps an input $\mathbf{x}$ through multiple non-linear transformations to an output prediction $\hat{y} = \arg\max_y p(y|x, \theta)$.

For different types of input data, there exist variants of this model that take advantage of the inherent structure of the data. Specifically for images, the main variant is a convolutional neural network (CNN) [27], with the key difference that the layer function is a convolution operation that only summarises local spatial information. This is enabled through weight sharing to keep the number of parameters manageable when scaling to high-dimensional images. There does not necessarily exist an analytical solution to the optimal model parameters θ as the parameter space is too big and non-convex. Therefore, training the model is done through stepwise optimization. In the research of this thesis, we have used the ADAM optimizer [28] to train CNNs with different composition of the number of layers and flow of information between the layers. There are many different architectures that are suitable for pathology and WSIs, where classification, detection and segmentation networks are considered a high-level categorization. Each of these have different variations, especially of the output layer, that let them model a problem better than the other, while still sharing some fundamental principles in the feature extraction. For more detailed description on deep neural networks for computational pathology, we refer the reader to Srinidhi et al. (2021) [29].

Chapter 3

Uncertainty and dataset shifts

In this chapter, we focus on aspects and risks that are relevant during a deployment setting of computational pathology, and how we potentially can mitigate these with current methods. We will give an introduction to uncertainty in deep neural networks, and some of the most popular approaches to quantifying it. Next, we will take a specific application in pathology, lymph node metastasis detection, and introduce some of the dataset shifts that appear "in-the-wild" in pathology under different deployment scenarios. Finally, we discuss the impact our contribution (c.f. **Appendix A**) has on future directions of these topics.

This chapter cites the first of the contributions of the thesis:

A. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J. D., & Dahl, A. B. (2020). Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *Proceedings of 23rd International Conference on Medical Image Computing and Computer Assisted Intervention* (pp. 824-833)

3.1 Uncertainty in neural networks

For safety-critical autonomous systems, we desire to have a quantitative measure of how much the system does not know about the problem. It will allow us to plan accordingly when we surpass a certain confidence threshold. This is the goal of uncertainty quantification (UQ), an active field of research because it is often remarked that neural networks fail to increase their uncertainty when predicting data far from the training distribution [30]. Whilst in everyday usage, uncertainty refers to being unsure, we can model it as the variation of y when drawn from a predictive distribution $p(y|x, \theta)$. Further, we can decompose it into two types for modeling purposes: aleatoric and epistemic [31, 30]. Aleatoric, also called data uncertainty, covers the uncertainty due to classes that overlap in input space, which originates from the data generation process. For example, a network trained on MNIST [1] should have aleatoric uncertainty if asked to classify an input appearing between '1' and '7' [32]. Epistemic refers to uncertainty about the model or its parameters, usually as a result of not collecting all possible data. E.g. if the same MNIST network is asked to classify an image of clothing, it should be uncertain due to a lack of knowledge about how to handle this type of input [32] because it was not in the data distribution. Therefore, this type of uncertainty is also sometimes referred to as model uncertainty.

While the concept of data and model uncertainty is well established from a theoretical point of view, we need to highlight that there are still no definitive answers in the literature on how these should be obtained in practice. In this thesis, we take a pragmatic approach and settle for estimating the predictive uncertainty as to the sum of the two types of uncertainty. In the next sections, we cover different popular approaches to obtain predictive uncertainty in deep neural networks.

3.1.1 Non-Baysian vs. Bayesian approaches

At a very high level, there are Bayesian and non-Bayesian methods, where one could argue that there are varying degrees of Bayesian as well. We mention the Bayesian approach here to easier relate the other methods, but it is not a focus in this thesis or **Appendix A** as these have not shown results that match other more heuristic methods on real-world applications, yet.

Bayesian neural networks [33] assume a prior distribution $p(\theta)$ over the network weights θ, and approximate the posterior distribution via the likelihood

[1] classical ML dataset of handwritten digits

function $p(y|\theta)$. Both coming up with a good prior, and performing inference to obtain the posterior of large models are fundamentally difficult. Even with advances in approximate inference [28], it is yet to be seen that the methods scale to large models in computational pathology.

Monte Carlo Dropout, originally proposed by Gal & Ghahramani (2016) [30], is a (rough) approximation of variational inference to obtain a distribution over functions. The basic idea is to take advantage of *Dropout* [34], an existing neural network regularization technique, to approximate the predictive uncertainty. Dropout formulates a simple layer operation, where a Bernoulli variable with a parameter p is element-wise multiplied with the output of the previous layer, hence randomly "dropping out" units. When these layers are added to the network during training, we avoid overfitting to the training data noise, i.e. obtain higher robustness to small perturbations. Now, remember that one forward-pass through a neural network is a point prediction of the function fitted with the network. To infer stochasticity into this framework, Gal & Ghahramani proposed to keep these Dropout layers during inference, and instead of making one single forward-pass, we make M stochastic forward passes. That is, for an input x', we can approximate the predictive mean

$$E_{q(y'|x',\theta)} = \frac{1}{M}\sum_{m=1}^{M} q_m(y'|x') \tag{3.1}$$

with variance as

$$Var_{q(y'|x',\theta)} = \tau^{-1} + \frac{1}{M}\sum_{m=1}^{M} (q_m(y'|x') - E_{q(y'|x',\theta)})^2 \tag{3.2}$$

where τ is a scaling hyperparameter for the model's precision on the data [30], typically found with cross-validation.

The method has obvious advantages as it re-purposes a layer already existing in many neural networks and allows us to use a single model during deployment. However, its drawbacks include hyperparameter tuning of p, τ, and also the optimal number and position of dropout layers in the network. All these are potentially very dataset dependent, and might not let the approach scale to so large models than originally published on.

Deep Ensembles are also an approximation to obtain the distribution over functions, yet more explicit and simpler than MC-Dropout. Lakshminarayanan et al. (2017) [35] popularized ensembles [36] for uncertainty quantification, and proposed to independently train M networks with the random initialization, letting each model m output a point prediction such that the predictive mean can be approximated as

$$E_{q(y'|x',\theta)} = \frac{1}{M}\sum_{m=1}^{M} q_m(y'|x') \tag{3.3}$$

with variance as

$$Var_{q(y'|x',\theta)} = \frac{1}{M}\sum_{m=1}^{M} (q_m(y'|x') - E_{q(y'|x',\theta)})^2 \tag{3.4}$$

where the final prediction is a simple averaging of M deterministic functions trained on the same data. Deep ensembles have clear benefits in their simplicity and parallelism opportunities - both from a training and deployment perspective. There is also empirical evidence that deep ensembles improve both predictive classification and quality of UQ [35]. The main drawback is the computational requirement for training M-independent models, and the theoretical understanding of why deep ensembles just trained random initialization work so well in practice. However, Fort et al. (2020) recently hypothesized it is due to how the method diversifies the function space within training trajectories [37]. By investigating the loss landscape of neural networks, i.e. the space of weights that the network navigates during training, they showed that deep ensembles explore different modes in function space. They also showed that so-called subspace sampling methods (e.g. Monte Carlo dropout) remain similar in function space, which produces an insufficiently diverse set of predictions. This gives some insights to understanding the dynamics of these methods, but there exist many different views of this subject.

3.1.2 Other views on uncertainty

A general aspect of many neural networks is the combination of a softmax output and cross-entropy loss for classification. For many papers, and ours included (Appendix **A**), softmax output is interpreted as model confidence and used as a baseline in experiments. The above methods are alternatives to this interpretation as it can have several pitfalls; the networks are generally overconfident [38], subject to manipulation by adversarial examples [39], and have issues with handling data outside the training distribution [30]. Nevertheless, it empirically performs moderately well as an indicator of predictive uncertainty. In brief, a softmax layer in isolation can learn to output a probability in between 0 and 1 to catch overlapping classes (aleatoric) but fails to decrease its confidence if queried far from the training data (epistemic) [32]. Whilst, softmax confidence remains an imperfect measure of uncertainty, recent experiments on standard benchmark tasks suggested final-layer feature overlap is more responsible for failures than softmax extrapolations [32]. In the work of this thesis, we have also looked at how the curvature of the loss landscape (i.e. smoothness), in addition to input and feature overlap impacts the quality of UQ.

Mixup was introduced as data augmentations technique [40], and have been empirically shown to improve predictive performance, and robustness to adversarial noise [40, 41, 42]. First, we recap the definition of Mixup. Consider a training set of input and output pairs $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \in \chi \subseteq \mathbb{R}^p$ and $y_i \in \chi \subseteq \mathbb{R}^m$ drawn from a joint distribution $\mathcal{P}_{x,y}$. During training, two random inputs (x_i, x_j) and their corresponding labels (y_i, y_j) are "mixed" together:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1 - \lambda) x_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j \end{aligned} \tag{3.5}$$

where $\lambda \in [0, 1]$ determines the mixing ratio of the linear interpolation. λ is commonly drawn from a symmetric Beta distribution $Beta(\alpha, \alpha)$ for $\alpha > 0$ controlling the strength of the input interpolation and the label smoothing. This means that the method generates new samples by linear interpolation of multiple samples and their labels. Recently, it has also been shown to introduce improvement in UQ, where networks trained with Mixup show less overconfidence [43]. The authors suggest that this is a consequence of training with smooth labels compared to hard labels. However, Zhang et al. (2021) show that Mixup augmentation is a data-adaptive regularization on the loss function, which can reduce overfitting and lead to better generalization behaviors than those of standard training [41].

Other (failed) approaches to UQ were also considered during this thesis. Studies have shown links between softmax unreliability and adversarial examples [39], we were inspired to look at methods that combat these vulnerabilities. We have not focused on adversarial robustness, but in brief, neural networks can be very sensitive to human imperceptible perturbations that easily flip output predictions [44], e.g. adversarial examples make an image of a panda be classified as a gibbon with high confidence. One approach that leads to adversarial robustness [45] studied a new regularizer that directly minimizes the curvature of the loss surface. They hypothesized that sharp changes in the geometry of the classification landscape and decision boundaries contribute to the small changes in input space leads to catastrophic changes in the output. They showed the existence of a strong correlation between small curvature and robustness and using second-order curvature regularization improved adversarial robustness. Therefore, we hypothesized that the same sharpness would restrain any "slack" on decision boundary between classes, hence only allowing the neural network to output high confident predictions. Our experiments (not shown in this thesis) indicated that such explicit regularization of the loss landscape indeed impacted UQ. However, enforcing high smoothness also acts as a strong regularization that significantly deteriorates the predictive performance - to such a degree that the performance is no longer relevant for any practical use in pathology.

3.1.3 Evaluation of uncertainty quantification

In this section, we recap different tasks and/or metrics that allow us to quantify any improvement that the methods contribute. In general, uncertainty is closely tied to calibration [38], but in this thesis, we are also interested in learning what it would practically influence or allow us to do when creating pathology applications. Therefore, we also use tasks, that if solvable by UQ improvements, would have great applicable benefits in practice.

Uncertainty calibration refers to the problem of predictive probability estimates that are representative of the true correctness likelihood [38]. Intuitively, if we consider a set of predictions that have average confidence of 60%, does this mean that we can expect 60% of the predictions to be correct? This is the notion of Expected Calibration Error (ECE) [46], a convenient summary statistic capturing the difference in expectation between accuracy and confidence. First, we compute the confidence of each of N observation denoted $p(\hat{y}_n)$, and bin these into H bins. We then calculate the ECE by comparing the content of each bin to its average accuracy. Let B_h be the set of indices for bin h. We calculate the bin accuracy

$$\mathrm{acc}(B_h) = |B_h|^{-1} \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \tag{3.6}$$

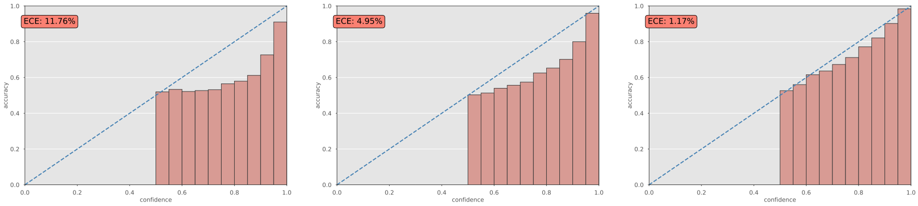
and the bin confidence

$$\mathrm{conf}(B_h) = |B_h|^{-1} \sum_{n \in B_h} p_n(\hat{y}) \ . \tag{3.7}$$

Finally, we calculate the weighted average of difference between the bins' accuracy and confidence:

$$\begin{aligned} \mathrm{ECE} &= \frac{1}{N} \sum_{h=1}^{H} |B_h| \cdot |\mathrm{acc}(B_h) - \mathrm{conf}(B_h)| && (3.8) \\ &= \frac{1}{N} \sum_{h=1}^{H} \left| \sum_{n \in B_h} p_n(y) - \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \right| && (3.9) \end{aligned}$$

where $\delta(x) = 1$ if $x = 0$ or $\delta(x) = 0$ if $x \neq 0$, and y_n^* is the true label. This metric allows us the quantify the calibration gap, where perfect calibration (complete agreement between accuracy and confidence) is zero and increasing ECE is a



(a) ECE = 11.76

(b) ECE = 4.95

(c) ECE = 1.17

Figure 3.1: Examples of three differently calibrated models from **Appendix A**. (a) shows an overconfident model, which leads to higher ECE. (b) shows a less overconfident model. (c) shows an almost perfectly calibrated model with ECE close to zero.

measure of higher miscalibration, see Figure 3.1. Together with standard predictive performance metrics, we could use this metric to quantify the impact of our experimental setup on calibration.

Misclassification detection involves downstream evaluation of uncertainty estimation on a specific problem and acts as a proxy task of UQ quality relevant for many real-world applications. As the name alludes to, it is a binary classification problem of detecting wrong predictions, hence Appendix **A** uses conventional metrics for binary classification. As pointed out by Ashukha et al. (2020), there are several challenges to use this to compare the quality of UQ between different methods. Every method induces its binary classification problem as the individual correct and incorrect predictions are model-specific, i.e. the dataset is not kept constant across methods, hence they solve different classification problems [47]. Even though it might be an imperfect measure between methods, we chose to keep it in Appendix **A** because it still provides valuable insights into one of the most relevant attributes from an automation perspective. If the model could pass relevant misclassified examples to a secondary backup (e.g. manual review) without burden it (e.g. passing all predictions to secondary), we would improve the safety of the system by avoiding potential catastrophic failures. At the same time, we also postulate that a system that has a high performance in misclassification detection would increase trust with the end-user (e.g. pathology department) when adopting automatic systems.

Out-of-distribution detection is another relevant task for pathology applications as neural networks can assign high-confidence predictions to inputs that did not belong to one of the training classes [39]. These examples are termed out-of-distribution (OOD) inputs. This behavior is not optimal in any high-stake application such as pathology because rare incidental findings, which are clinically relevant, may then be missed by an algorithm because they are outside the distribution of the training set. For deep learning in general, there has been a decent amount of research on this topic [48, 49, 50, 51, 41]. But as pointed out by Winkens et al. (2020) [52], the difficulty level depends on how semantically close the OOD inputs are to the training classes. They propose to distinguish between far-OOD (easy) and near-OOD tasks (harder). As with most research in machine learning, OOD research has been driven by benchmark datasets. UQ has used mostly far-OOD tasks, e.g. training on MNIST and detecting SVHM[2] as OOD inputs. Training on CIFAR10[3] and detecting novel classes from CIFAR100[4] as OOD inputs is an example of a near-OOD task. In the research thesis, we wanted to study a near-OOD case as this is clinically more relevant, and it is a realistic task considering the dataset shifts that occur in pathology.

3.2 Dataset shifts in pathology

Dataset shifts occur when the input and/or output distributions differ between what neural network was trained on and what is seen during testing or deployment. It is a common problem present in most predictive modeling applications due to many different reasons ranging from pre-analytical errors to uncalibrated cameras and potential (un)intentional biases in the training data. Because of the high complexity of pathology, we postulate that dataset shifts are more frequent and diverse than in most other fields and are probably the single most contributing factor for the lack of generalization. It is mainly due to the natural biological variance of cellular level data, but also due to the lack of standardization, especially on pre-analytical and analytical variables (see section 2.2.1).

The hope is that well-designed systems will alarm when a significant dataset shift occurs. However, as previously mentioned, neural networks tend to fail silently, and unfortunately, in practice, machine learning pipelines rarely inspect incoming data for signs of distribution shifts. The uncertainty methods described in the previous section are all attempts to embed this desired property into the UQ of the model itself. For this thesis, we aimed to investigate relevant dataset shifts in pathology that were as realistic as possible for clinical applications. Therefore, we used a concrete example application.

[2] Dataset of house street numbers

[3] Dataset of 10 different classes such as airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

[4] Similar to CIFAR10 but with 100 distinct classes

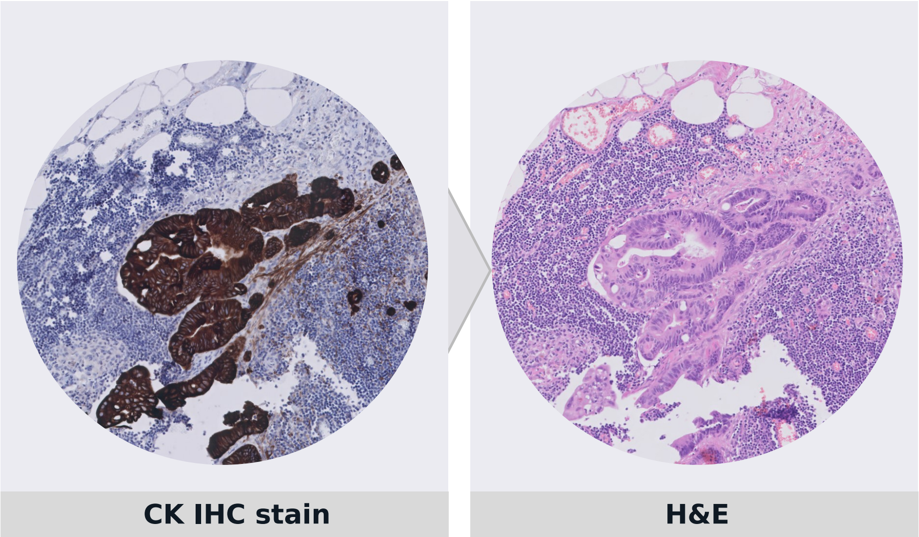


Figure 3.2: Example of a lymph node metastasis in a lymph node section. Left shows it stained with cytokeratin (CK). Right shows the same metastasis in a H&E stained section.

3.2.1 Lymph node metastases detection

The presence of lymph node metastases is an important prognostic factor for cancer patients and is integrated into the TNM-diagnostic grading system (section 2.1.1). The manual examination process is time-consuming and can lead to small metastases being missed. It is probably one of the most known problems within pathology from a machine learning perspective due to the CAMELYON challenge [53, 54] - an ImageNet-like impact on the field of computational pathology. In brief, the aim was to develop systems for the detection of breast cancer metastases in lymph nodes, and the organizers released one of the largest labeled WSI datasets [55]. Since then, multiple research studies [56, 57, 58] have shown promise and commercially available algorithms have either been clinically approved as a medical device [59] or are under research-use-only development [60]. Taking all of this into account, this application was an obvious choice to study because it allowed us to focus on unknown aspects outside the regular predictive performance. Moreover, we had a unique change of utilizing public and private datasets to create a controlled real-world experimental setup.

3.2.2 In-distribution dataset shifts

We define in-distribution shifts as valid changes in the input distributions that primarily originate from the data generation process under different settings. This involves any change in pre-analytical steps such as fixation and sectioning, e.g. if the sectioning thickness is modified. Other common changes are related to analytical steps such as staining, especially variability in H&E staining, but also different scanners impact the appearance of the input data [61]. We decided to capture some of these in a realistic development setup, where a dataset shift might happen when going from the training data to the test data of different origins. For any clinical pathology application, we usually perform internal and external validation. Internal refers to testing on data originating from the same hospitals as the training data, and external refers to testing on data generated on a different hospital. These testing schemes are usually to validate the generalizability of the predictive performance. External testing allows us to investigate what happens in a valid product deployment scenario (not to confuse with prospective clinical validation).

Other changes in the data distribution are inherent to pathology. Cancer histology is a complex classification of tumors based on the type of cells in which cancer originates (histological type). For simplification, we only focus on the most common type; malignancies of epithelial tissue called carcinoma. This type refers to a malignant neoplasm of epithelial origin or cancer of the internal or external lining of the body. Carcinomas account for 80 to 90 percent of all cancer cases [62]. Carcinomas are divided into two major subtypes: adenocarcinoma and squamous cell carcinoma, which originates in a gland, and the squamous epithelium, respectively. Most carcinomas affect organs or glands capable of secretion, such as the breasts, which produce milk, or the lungs, which secrete mucus, or colon or prostate or bladder. Because the same cancer subtype can originate from different organs and metastasize to lymph nodes regardless of origin, we collected lymph node sections with adenocarcinoma from colon cancer. Hence, by training a model to recognize adenocarcinoma from breast cancer on lymph node sections, and testing on adenocarcinoma from colon cancer, we keep the fundamental biological factors of the input and output distribution fixed, but we implicitly introduce a small shift in the input distribution.

3.2.3 Out-of-distribution dataset shifts

Similarly, we define out-of-distribution shifts as unexpected changes in the joint distribution of inputs and outputs making them fundamentally different than the training distribution. Especially, near-OOD inputs are relevant to study in pathology because it is characterized by large amounts of possible anomalies. Making sure that these do not significantly impact the intended use of a trained (specialized) model is crucial for successful implementation into practice. Some rare findings might be clinically relevant to report (e.g. other malignancies), where others such as scanning artifacts simply need to be ignored.

To study this as realistically as possible, we collected lymph node sections with squamous cell carcinoma from head and neck cancer. This can be seen as near-OOD inputs as it is still cancerous cells, however, the histo-morphological characteristics are different than adenocarcinoma, i.e. they do not necessarily look the same. To make it even more difficult, the subtypes are also graded ranging from well- to poorly differentiated. The morphology of well-differentiated cancer cells is more like the normal cells in the tissue they started to grow in, e.g. a well-differentiated adenocarcinoma will look similar to normal epithelial cells. Poorly or undifferentiated cancer cells look very morphological different from their origin normal cells. These cells look immature, undeveloped, or irregular and are not organized in the same pattern as normal cells [4]. Moderately differentiated cancer cells look and behave somewhere between well- and undifferentiated cancer cells. When taking these considerations into account, we can look at the morphology variability between adeno- and squamous cell carcinoma (SCC) as a spectrum (see Figure 3.3 graphical representation with examples). This makes our experiments in Appendix **A** interesting because we mimic a day-to-day task that a pathologist would and can perform. Concurrently with our work, Linmans et al. (2020) [63] proposed a similar experiment with diffuse large B-cell lymphoma, which is a different tumor classification than carcinoma. This can also be deemed a clinically relevant near-OOD task. They propose a more computationally efficient version of an ensemble and find similar results to our findings in **Appendix A**. In general, many applications in pathology will have near-OOD inputs, where fine-grained details in cell and tissue patterns introduce dataset shifts. It only enforces the need to study this phenomenon in computational pathology.

3.3 Experimental findings (Appendix A)

In this section, we provide a high-level summary of the experimental findings in **Appendix A**. All current state-of-the-art methods can generalize in terms of predictive accuracy from the internal test set to the external dataset with only a small impact on the calibration of the predictive uncertainty. When introducing near-OOD inputs, all investigated methods show both decreased performance and higher overconfidence. For the in-distribution dataset shift, we found a similar behavior even though the performance decrease was smaller than under

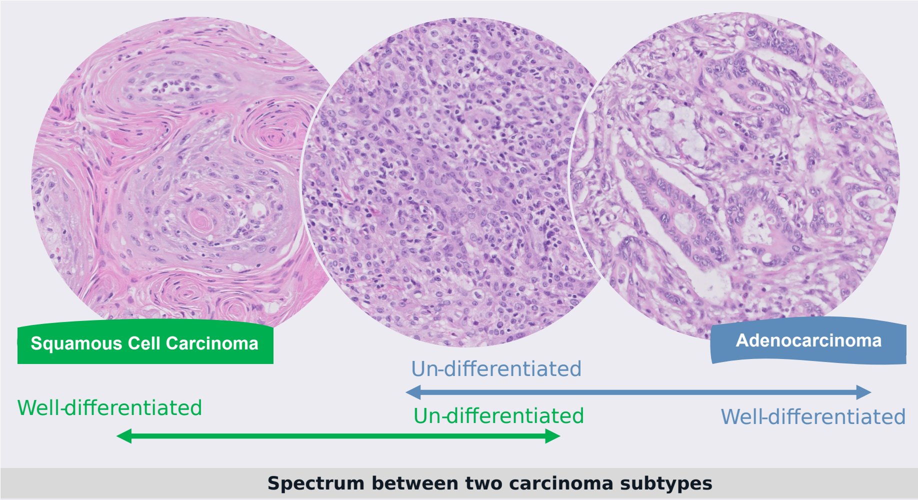


Figure 3.3: Examples of morphological differences between adenocarcinoma (blue) and squamous cell carcinoma (SCC) (green). Adenocarcinoma has tendency to grow in glandular structures, whereas SCC have a swirling pattern, but when both become poorly differentiated, they loose their original characteristics (middle).

the out-of-distribution shift. Our experiments also showed minimal benefits of two of the methods intended to improve UQ compared to the baseline method, and MC-Dropout can potentially hurt the calibration performance on all dataset shifts. We refer to **Appendix A** for the full detailed description of our findings.

3.4 Concluding remarks

Our main contribution to the topic of uncertainty and dataset shifts in pathology is a thorough investigation of several state-of-the-art methods' ability to quantify uncertainty while keeping high accuracy. This has only been covered in previous investigations on popular benchmark datasets of natural images [64, 65]. However, this is insufficient for pathology evaluation because the appearance and variation resulting from distributional shifts of histopathology images are very different from those of natural images. Therefore, we extended our evaluation to a unique multi-hospital single indication training set and performed an extensive evaluation on both internal and external test sets and clinically plausible distributional shifts. We believe that our contribution has shed new lights on how we can evaluate existing and future UQ methods in a realistic real-world pathology setting.

There still remains a lot of recent research not included in this thesis such as investigating the generative models [66, 28], contrastive learning [67, 52], and newer architectures [68, 37] in relation to capturing models' uncertainty and ability to use this on practical tasks. Similar, an open question is how the uncertainty component impacts fields like active learning [69], where we aim to collect more data in unexplored and uncertain regions to lower the manual annotation burden while improving the predictive performance of the model. There is still much work to be done to understand how different dataset shifts in pathology affect a deep neural network, and how we use that information beneficially, e.g. by mitigating the risks of silent failures. Here, normalization and augmentation schemes seem to handle in-distribution shifts such as minor staining and scanning variability in H&E stain. However, one aspect not covered in this thesis is how to monitor and adjust for adequate staining on IHC biomarkers, where a normalization might introduce false-positive biological signals or the expression is not reflecting the true signal due to staining procedures. All the research in this chapter aims to get the predictive model to also handle out-of-distribution shifts but other alternative approaches should be explored, such as upstream monitoring of dataset shifts.

Chapter 4

Prognosis of cancer by quantifying the immune system

In the previous chapter, we introduced and investigated the influence of certain dataset shifts that can happen when deploying an algorithm for a well-established pathology application. In this chapter, we take a look at what it takes to develop an algorithm for a new emerging biomarker, and what pitfalls and challenges are associated with doing so. There are many biomarker candidates, but this thesis focused on tumor infiltrating lymphocytes (TILs) for TNBC patients for several reasons: (i) It is associated with prognostic and predictive capabilities for TNBC and HER2+ patients. (ii) Its hierarchical complexity allows us to identify, and document the experience of adopting a guideline for implementation. (iii) It is a biomarker not yet fully implemented in the clinic, hence pathologists' could be more apt to consider it than asking them to apply an ML tool for a biomarker that they have assessed differently for decades already. First, we will give an introduction to the requirements of following a pathology guideline and the variability of data used in **Appendix B**. Secondly, we describe some of the deep learning methods developed to comply with the guideline, and how biomarker quantification is used for survival analysis of patients. Last, we will discuss some of the challenges and pitfalls (c.f. **Appendix C**) associated with ML for this specific biomarker, and what the future perspectives of our findings have for TILs and other similar computational biomarkers.

This chapter cites the second and third contributions of the thesis:

B. Thagaard, J., Stovgaard, E.S., Vognsen, L.G., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Vincentz, R.E., Jepsen, R.K., Roslind, A., Kümler, I., Nielsen, D., & Balslev, E. (2021) Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers. In *Cancers* 13(12):3050.

C. Thagaard, J., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Roslind, A., Nielsen, D., Balslev, E., Salgado, R., ..., & Stovgaard, E.S. (2021) Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer. To be submitted.

4.1 Pathology guidelines

In this section, we briefly recap the purpose of guidelines, the basics of the current international guideline on TIL assessment in breast cancer, and how these can also help in the development of computationally approaches.

Pathology is an inherent complex medical discipline that requires extensive training of medical practitioners to even specialize in a subarea, e.g. breast cancer. Guidelines are usually specific to a certain combination of indication and biomarker and offer guidance on how to interpret the combination in clinical practice. They are often created by regional or national organizations (e.g., American Society of Clinical Oncology (ASCO), College of American Pathologists (CAP), European Society for Medical Oncology (ESMO), etc.) or self-organizing working groups of pathologists (e.g., International Immuno-Oncology Biomarker Working Group on Breast Cancer (TIL-WG)). Adopting these evidence-based guidelines help pathologists and other clinicians to make more informed and standardized decisions about diagnosis and optimal treatment for the patients.

In this thesis, we focus on the guideline proposed by TIL-WG [70] but many of the general considerations can be applied to other indications. The purpose of the guideline is to answer the following questions: (i) why are TILs clinically important and (ii) how to score TILs manually? In section 2.1.2, we summarized the answer to the first question. Interestingly, the answer to the second is a step-by-step manual guide - an algorithm - with inclusion and exclusion criteria for assessing TILs in breast cancer [70]. Briefly, it states to distinguish between intratumoral TILs (iTILs) in direct contact with tumor cells, and stromal TILs (sTILs), which are located in the stromal tissue between islands of tumor cells. The recommendation is to focus on sTILs, as evaluation of these is more reproducible [19, 71]. sTILs are then assessed as a percentage area coverage of total stromal tumor area and reported as a continuous variable (see Table 4.1 for all steps). Even though standardization and training efforts have been shown to increase the reproducibility of manual scoring [18, 19, 72, 73], there exist inherent pitfalls that could hinder the implementation into the routine clinical management of breast cancer [74]. Therefore, there has been an expectation of the promise and potential of automated image analysis to overcome some of the limitations of visual TIL assessment (VTA) [75]. The TIL-WG has been actively working with the ML community and even produced a report on how computational assessment of TILs could be designed [76]. The recommendation of this work is that; *"computational TIL assessment (CTA) algorithms need to account for the complexity involved in TIL-scoring procedures, and to closely follow guidelines for visual assessment where appropriate"* [76]. However, all existing studies on CTA have proposed alternative quantitative metrics (e.g., lymphocyte percentage, spatial patterns of lymphocyte distribution) for sTILs assessment rather than being consistent with the guideline. Also, they all involved some aspect of manual work (e.g. define the tumor region, exclusion of DCIS), hence still prone to intra- and interobserver variability while not being suited for workflow 3 implementation (section 2.2.2). These aspects are key to our contribution in **Appendix B**.

Because the guidelines closely resembles algorithmic-steps, they serve well as a recipe to develop computational approaches. It does require some interpretation, e.g. in step 8 (Table 4.1), the definition of a hotspot is not specified quantitatively. Due to the guideline's hierarchical structure with both tissue- and cell-level requirements, it requires some specific design considerations and methods to recognize some of the tissue and cells classes relevant for TILs in TNBC, see Figure 4.1 for an overview.

4.2 Different problems require different methods

In this section, we introduce the software platform used in **Appendix B** and describe the fundamental methods used to covert the guideline into a computational algorithm.

1. TILs should be reported for the stromal compartment (=% stromal TILs). The denominator used to determine the % stromal TILs is the area of stromal tissue (i.e. area occupied by mononuclear inflammatory cells over the total intratumoral stromal area), not the number of stromal cells (i.e. fraction of total stromal nuclei that represent mononuclear inflammatory cell nuclei).
2. TILs should be evaluated within the borders of the invasive tumor.
3. Exclude TILs outside of the tumor border and around DCIS and normal lobules.
4. Exclude TILs in tumor zones with crush artifacts, necrosis, regressive hyalinization as well as in the previous core biopsy site.
5. All mononuclear cells (including lymphocytes and plasma cells) should be scored, but polymorphonuclear leukocytes are excluded.
6. One section (4–5 μm, magnification ×200–400) per patient is currently considered to be sufficient.
7. Full sections are preferred over biopsies whenever possible. Cores can be used in the pretherapeutic neoadjuvant setting; currently, no validated methodology has been developed to score TILs after neoadjuvant treatment.
8. A full assessment of average TILs in the tumor area by the pathologist should be used. Do not focus on hotspots.
9. The working group's consensus is that TILs may provide more biologically relevant information when scored as a continuous variable since this will allow more accurate statistical analyses, which can later be categorized around different thresholds. However, in daily practice, most pathologists will rarely report for example 13.5% and will round up to the nearest 5%–10%, in this example thus 15%. The pathologist should report their scores in as much detail as the pathologist feels comfortable with.
10. TILs should be assessed as a continuous parameter. The percentage of stromal TILs is a semiquantitative parameter for this assessment, for example, 80% stromal TILs means that 80% of the stromal area shows a dense mononuclear infiltrate. For assessment of percentage values, the dissociated growth pattern of lymphocytes needs to be taken into account. Lymphocytes typically do not form solid cellular aggregates; therefore, the designation '100% stromal TILs' would still allow some empty tissue space between the individual lymphocytes.
11. No formal recommendation for a clinically relevant TIL threshold(s) can be given at this stage. The consensus was that a valid methodology is currently more important than issues of thresholds for clinical use, which will be determined once a solid methodology is in place. Lymphocyte predominant breast cancer can be used as a descriptive term for tumors that contain 'more lymphocytes than tumor cells. However, the thresholds vary between 50% and 60% stromal lymphocytes.

Table 4.1: Recommendations for manual assessing TILs in breast cancer from Salgado et al., (2014) [70]

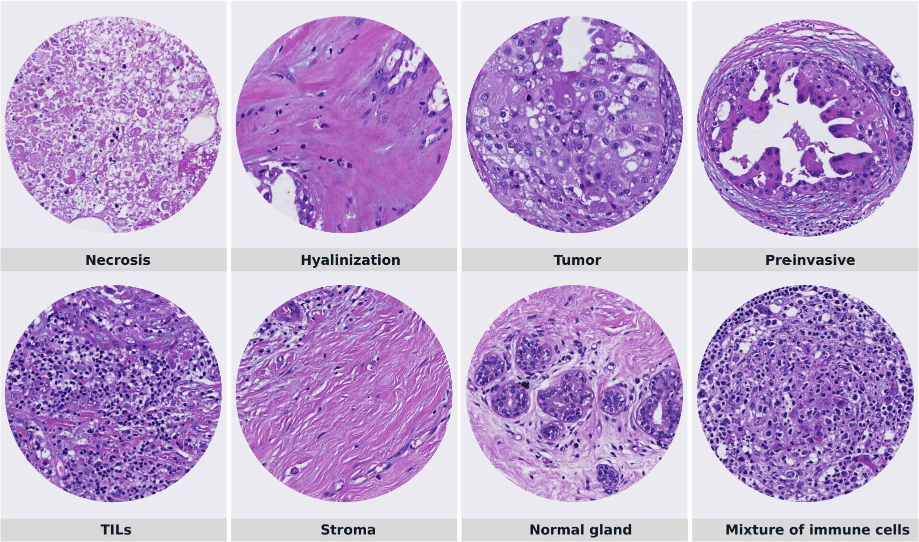


Figure 4.1: Example images of the variability of structures which tissue- and cell-level model need to comprehend.

4.2.1 Visiopharm AI platform

The Visiopharm AI platform is a general-purpose image analysis platform, specifically tailored to create and deploy algorithms for pathology images. The platform uses general concepts that enable us to go from a digital image to quantitative output results. These steps include [77]:

1. **Preprocessing** includes aspects such as noise removal, image normalization, etc. where the image information is not changed significantly.
2. **Feature engineering** enhance certain signals in the image, e.g. a blob-filter enhances the signal for round objects in the image, or image deconvolution, where the DAB signal in the image is isolated from the hematoxylin.
3. **Classification** assign a certain class to each entity (pixels, objects, or entire field-of-view), using (learned) rules that go from the input (pre-processed image or feature image) to the output, e.g. segmentation.
4. **Post-processing** is a powerful step in any image analysis algorithm, where the output of the classification can be further processed, and object-level heuristic and objective rules can easily be incorporated into the anal-

ysis pipeline, e.g. removing small nuclei below a certain area or other histology relevant rules that apply to the problem at hand.

5. **Output calculation** is usually the last step of any image analysis algorithm as until now we have not quantified the objects from the classification and post-processing steps. Most importantly, these are usually human interpretable formulas e.g. counting of TILs, measuring the area of stroma area, or the combination to obtain the density of TILs in the stroma as a number per mm^2.

These concepts are written here as high-level as possible as most image analysis algorithms can be described by either one or more combinations of these. Most importantly, we have in this thesis used deep neural networks to combine the feature engineering and classification concepts, replacing most of the difficult and tedious work of translating rules into the computer while still being complemented by the rest of the concepts. However, as Oscar Wilde writes, *"the truth is rarely pure and never simple"* ("The Importance of Being Earnest", 1895) because most real-life applications of image analysis in pathology exist across several conceptual levels. In practical terms that means that we have used both deep neural networks alongside more classical rule-based approaches to arrive at the most efficient and robust algorithm (c.f. **Appendix B**; **Appendix C**).

4.2.2 Differences in model architectures are important for histology

In **Appendix B**, we proposed to use two different deep neural network architectures to create a CTA algorithm. In this section, we recap the differences of these, for which problems they are well suited for, and why. We could have chosen to just use one type but initial experiments showed that it was beneficial to use different models. The main reason for that is how well the network architecture suited the tasks they should learn to solve.

U-Net is a popular CNN architecture in medical image analysis, originally proposed to segment cells in electron microscopic images [78]. The core structure is a contracting (encoder) and upsampling (decoder) neural network. The contracting pathway learns to extract and compress a feature representation of the image, whereas the upsampling, in a step-wise manner, learns to propagate low-level features in the bottleneck (deepest part of the encoder before decoder) with the contracting path information until the features are increased to the resolution of the input. In **Appendix B**, we chose this model for cell-level problems as it generates more precise outputs with fewer training images than other alternatives [78]. We also found that by using step-wise down- and upscaling, we were able to use more simple and easier obtainable training data (covered in section 4.3) which information would otherwise be lost in more aggressive sampling.

DeepLabV3+ [79] is also an encoder-decoder CNN architecture with one distinct difference compared to the U-Net; an Atrous Spatial Pyramid Pooling (ASPP) block in the bottleneck. The ASPP introduces a parallel multi-scale feature extractor across the entire spatial dimensions of the activation maps at the end of the encoder. The multi-scaling aspect is obtained by having parallel branches of a global average pooling (GAP) layer, and 4 different dilated/atrous convolution layers. The intuition behind the GAP layer is to capture features representing the entire image, e.g. a network trained to segment tumor cells gets an input with no tumor cells present, then it can use this feature to enforce that no other branch should be used to output any tumor cell segmentation. The purpose of the dilated convolution branches is to create features that represent different scales in the input image [80]. These two attributes make the DeepLabV3+ architecture well suited for tissue-level recognition and segmentation as it can use all contextual information in the input image at different scales without having many down sampling layers.

Several studies [81, 82] have shown that U-Net surpasses object detection architectures for the detection of lymphocytes in IHC and IF images. Similarly, DeepLabV+3 has been shown to outperform U-Nets on tissue-level segmentation tasks, e.g. for lymph node metastases [60], gastric cancer [83], and dermatitis [84]. Could we have selected different models, and created the same algorithm as in **Appendix B**? Probably, but these architectures are selected due that U-Net and DeepLabV3+ seem to fit modeling cell- and tissue-level patterns in pathology images.

4.2.3 Rule-based methods to infer object logic

Deep neural networks should be used to what it is good at and not everything. It is very difficult to make handcrafted robust features for tasks such as recognizing and classifying different cell types or segmentation of highly heterogeneous tumor regions. Here, it makes sense to use the predictive power of deep neural networks. However, there are also tasks where less complex methods are more suitable. For example, in step 2 (Table 4.1), to define the "borders" of the invasive tumor, we proposed in **Appendix B** to use morphological operations such as closing and opening [85] with a fitting kernel size on the segmented invasive tumor objects. It was implemented as post-processing steps and can be seen as a simple version of a closed concave hull algorithm [86] on all the invasive tumor cell nests, see Figure 4.2.

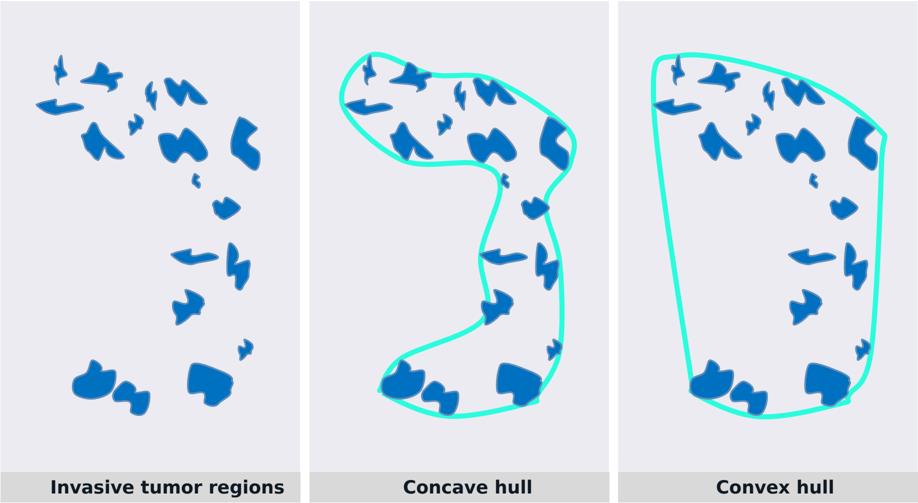


Figure 4.2: Illustration of the approximation of the concave hull using post-processing on the segmented invasive tumor regions. This creates a smooth macro-tumor outline around all the tumor nests similar to what a pathologist would define as the tumor border. In comparison to the convex hull, which would include too much stroma not associated with the tumor regions.

Another example is step 3 (Table 4.1), where TILs outside the tumor border and around DCIS and normal glands should be excluded. In **Appendix B**, we obtained the invasive tumor, DCIS, and normal glands as segmented objects, hence we could easily create a margin zone around these objects by using a distance-based rule via dilation, where we should not include TILs, see Figure 4.3. We refer the reader to **Appendix B** for a full description of the entire algorithm.

Even though the most difficult image analysis tasks are moved to deep neural networks, it does not come for free. Most of the work of creating an image analysis algorithm is now spend creating the training dataset, which can be a challenging task as well.

4.3 Obtaining objective training data

A key focus of this thesis was to investigate methods to create better training data with less effort for pathologists. We quickly discovered that creating manual pathologist training labels on H&E for the cell- and tissue-level CNNs

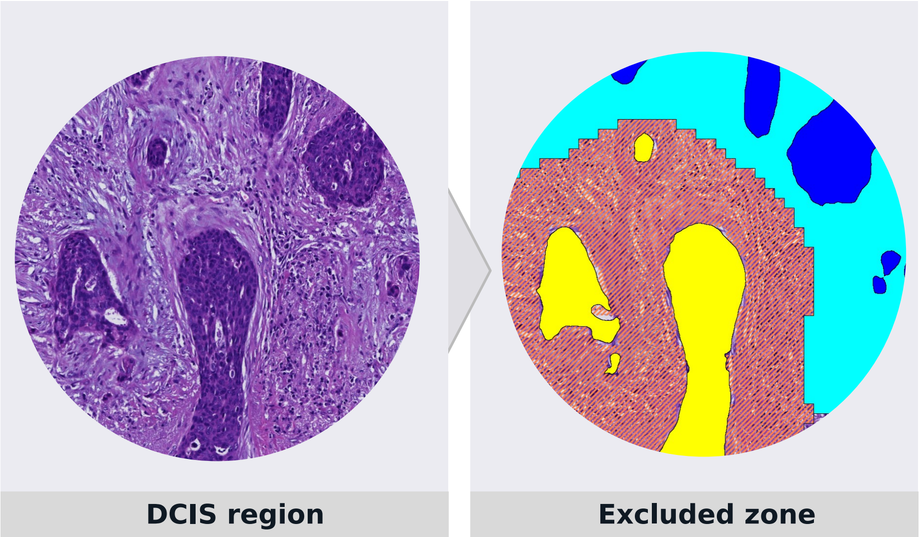
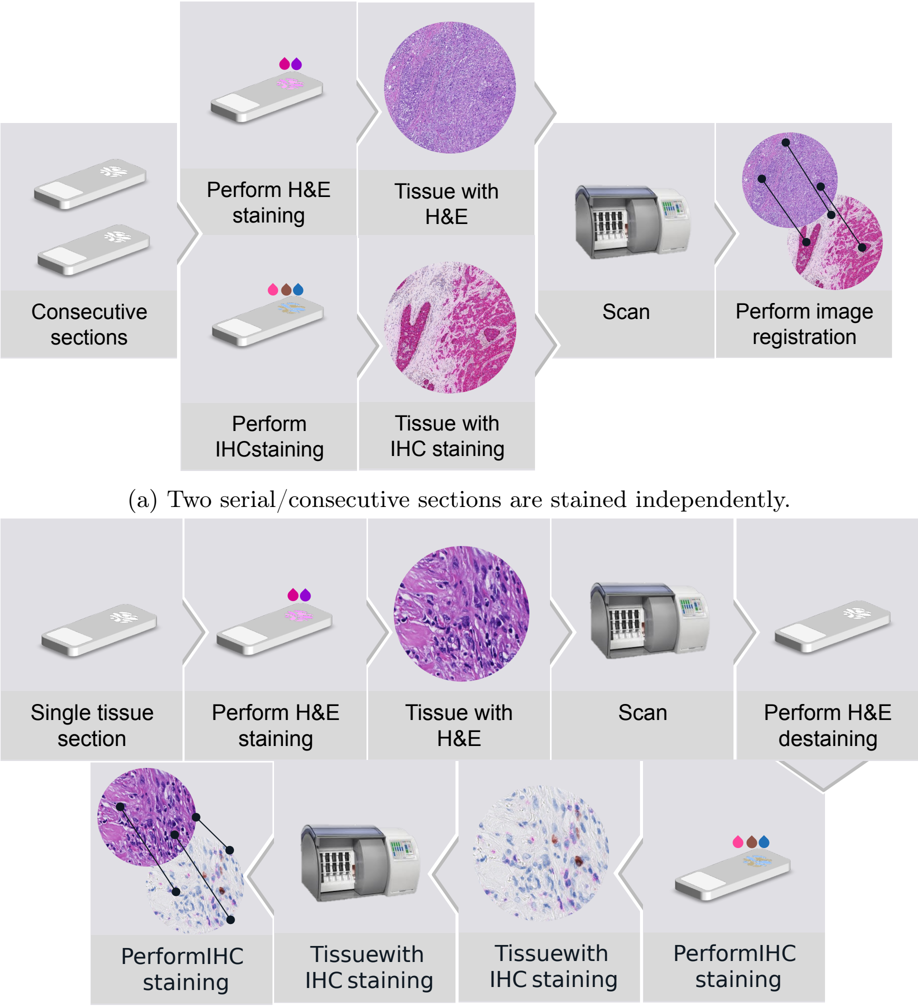


Figure 4.3: Example of how simple dilation from detected DCIS objects (yelllow) can create a zone around them (orange) where all TILs are excluded. This allow us to only include TILs stromal region (cyan) associated with invasive tumor nests (blue).

would not be sufficient to train models with high enough performance on these tasks. The two main reasons were pathologists' time (cost), and inconsistencies in manual labels for TILs detection. Therefore, we early on proposed two different strategies to obtain more consistent labels that only required a pathologist to review the final training data, namely (i) serial section or (ii) stain re-staining. The principle for both is to utilize an IHC-based strategy that also produces more objective ground truth compared to manual labels. This possibility is unique to pathology. The only, but important, difference between the two methods is either use of a serial section or the reuse of the same section, see Figure 4.4. For tissue-level tasks, serial sections are sufficient because epithelial cells are relatively large so the general structures are likely to be present between the two sections, and the shift is easily compensated for during registration. However, for the cell-level task, lymphocytes are so small that there is a high risk that one cell present in the first section is not present in the next. The stain re-staining approach allowed us to generate objective cell-level data for the two cell types defining TILs (see step 5 in Table 4.1).

Similar approaches have been used in other pathology studies [87, 88], however, one of our main contributions in **Appendix B** is to use physical double staining (PDS) to obtain epithelial regions (cytokeratin positive) and discriminating be-



(a) Two serial/consecutive sections are stained independently.

(b) The same section is stained, destained and re-stained.

Figure 4.4: Examples of the two different IHC schemes used in **Appendix B**. (a) shows the serial section approach for tissue-level label generation. (b) shows the stain re-staining methodology for cell-level label generation.

tween invasive and DCIS and normal lobules (loss of P63 positive myoepithelial cells) for the tissue-level labels. Similar, we could transfer both lymphocytes (CD3 positive) and plasma cells (CD79a positive) objectively to the primary H&E. Generally, we showed that IHC stains can be used during development to help guide semi-automatic labels transfer onto the H&E slide, which means that the models then can be trained and deployed on H&E only - one of the requirements of the VTA guideline.

4.4 Survival analysis

As recently discussed by Arcs et al. (2021), it is still an open question how to best validate CTA algorithms [89]. In **Appendix B**, we sanity-checked the concordance between manual TILs-scores and the automated approach but used the outcome of the patients to investigate the clinical impact. In this section, we describe the concepts of survival analysis in more detail.

The primary endpoint for assessing a biomarker is to associate the score of each patient to the time of an event of interest [90] with time to *overall survival (OS)* and *relapse-free survival (RFS)* commonly used as the events of interest. For OS and RFS, the definition of time is from surgery until the death of any cause, and until local or distant relapse of disease, respectively. An important term is *censoring* that captures if not all included individuals in the study who have experienced the event. The last visit date can be used to capture if a patient was lost in follow-up during the study period, but also death accounts as censoring for RFS. To study these endpoints, we used two standard statistics; (i) Kaplan-Meier survival estimate [91], and (ii) Cox proportional hazards (PH) regression analysis [92].

Kaplan-Meier survival estimate is a non-parametric model of survival probability $S(t)$ usually visualized in survival curves, where $S(t)$ is plotted against time t for different patient stratification. Let $t_1 < t_2 < ... < t_k$ be independent events (e.g. death) of k patients, then:

$$S(t_i) = S(t_{i-1}) \left(1 - \frac{D(t_i)}{N(t_i)}\right) \tag{4.1}$$

where $t_0 = 0$ and $S(0) = 1$ with D and N being the number of events and number of patients alive at certain time-point t_i, respectively [90]. This results in a step function that changes between events every time a patient experience an event.

Cox PH regression analysis is a simple statistical non-parametric regression model that allows us to investigate the effect of several clinical factors' association with an event simultaneously. The risk of an event at time t is modelled by the hazard function:

$$h(t) = h_0(t) \times e^{b_1 x_1 + b_2 x_2 + ... + b_p x_p} \tag{4.2}$$

where h_0 is the baseline hazard and $\{x_1, x_2, ..., x_p\}$ is a set of p variables (the biomarkers) [93]. From this, we can obtain the hazard ratios (HR) as e^{b_i} that allow us to interpret if a biomarker is positively or negatively associated with an event. A biomarker can either have no effect (HR=1), increased (HR>1), or reduced (HR<1) risk of an event. For example, if HR=1.2 of a specific biomarker and the event is death, it means that for every unit increase of the biomarker value, there is a 20% increased risk of dying. The Cox model is practically a multiple linear regression of the log of the hazard on the variables x_i, with h_0 being an 'intercept' term that varies with time [93].

4.5 Experimental findings (Appendix B & C)

In this section, we provide a high-level summary of the experimental findings. We demonstrated that CTA can provide a quantitative and interpretable score that correlates with the manual pathologist-derived sTIL status. However, a more influential finding, is that a CTA score can be prognostic for OS in patients with TNBC (HR: 0.81 CI: 0.72-0.92 p=0.001) independent of age, tumor size, nodal status, and tumor type in statistical modeling. We also showed that the quality and consistency of the labels generated with IHC labeling schemes were higher than manual annotations. Lastly, CTA was found to address some of the challenges of VTA such as reproducibility, but can suffer from pitfalls, e.g. challenges with technical factors and achieving high enough generalization to all variability seen in pathology. However, we also layout that many of these challenges can be solved as long as the right model, training data, and validation considerations are taken into account. For the full detailed description of our findings, we refer the reader to **Appendix B** and **Appendix C**.

4.6 Concluding remarks

Our main contribution is a complete development effort to adapt a manual guideline for assessing TILs in breast cancer into an automated method. We studied and documented a thorough strategy to create objective training data, enabling scaling of the process that usually is the largest obstruction for computational pathology. We collected and digitized a large retrospective dataset on which we successfully showed that a fully automated workflow 3 algorithm for TIL assessment is associated with overall survival, confirming the prognostic potential of TILs for TNBC patients. In **Appendix C**, as a follow up on **Appendix B**, we identified several challenges and pitfalls that can impact the performance, generalizability and cause discrepancies on outcome estimates when transferring a manual VTA guideline into a CTA algorithm. We hope our contributions highlight the potential and pitfalls in using machine learning for TIL assessment, and future studies will be armed to find the answers needed to ensure reliable and reproducible CTA into the routine clinical management of breast cancer.

There remains a lot of work to further develop and validate CTA in clinical studies [89]. Here, the insights from Chapter 3 should be included and studied further. With recent research not included in this thesis, there is also potential to study TILs further, such in combination with other biomarkers (e.g. tumor-stroma ratio [94]) or its predictive potential toward immunotherapy [95]. Another avenue to pursue is to discover information not possible for a pathologist to quantify, e.g. using recently advanced in geometric deep learning [96] to interrogate the heterogeneity and spatial distribution of TILs, or use all information to train a model to directly predict prognosis [97, 98].

Chapter 5

General discussion

In the work of this thesis, we explored several aspects of bringing deep learning-based algorithms into real-world settings. In this chapter, we discuss our results and conclusions in the larger context of the aim of this thesis, before we look at what we don't know yet as the foundation for providing future perspectives of this thesis.

5.1 Studying the impact of dataset shifts when deploying an algorithm into clinical practice

From a deployment and safety perspective, it is highly valuable to be able to report what the system can and cannot handle with high confidence. Therefore, in the early days of this PhD project, our expectation was high in terms of the state-of-the-art methods' ability to obtain reliable uncertainty estimations in deep neural networks. We identified several major gaps from the traditional studies in the field to the world of pathology. First, we noticed that the tasks were not representative of the difficulty of pathology as most benchmark datasets were so-called far-OOD inputs. Therefore, we set out to create representative tasks in pathology, which, if solvable, would bring immediate value to the diagnostic algorithms for detecting lymph node metastases. Secondly, the current research had only started scaling these methods to larger models coping with natural images, so we also wanted to investigate if these methods also apply to the appearance of pathology images.

Our research shows that the investigated methods can only provide reliable uncertainty estimates if used within the same data distribution as included in the training set, but one should not expect current methods to alarm novel abnormalities or all error cases. This means that currently, other steps need to be taken to ensure that incoming data is inspected for signs of distribution shifts. Here, interoperability to laboratory information systems (LIS) will become critical such that metadata can be verified. Good laboratory practices should ensure that processes and controls are in place to mitigate several of these datasets shift after the initial validation of the algorithm, with a stated goal of standardizing the diagnostic process. These include quality systems, the use of controls, continuous training, and enrolment in external quality assessment (EQA) programs. If such systems are implemented, we see that algorithms can be adopted into the workflow as an assistance tool with pathologists reviewing the results as a minimum. This means that fully autonomous algorithms are still years out in pathology and the diagnostic responsibility should continue to reside with the pathologist.

While we did not investigate the impact of dataset shift on the TIL application, we expect our findings to also apply to this and other applications. With the vast number of different use-cases in pathology, there is still a need to understand the benefits and pitfalls of every single one to ensure that the predictive ML tools can be deployed safely and risk of catastrophic failures can be mitigated. Therefore, we do not believe that new methods that monitor dataset drifts should necessarily be integrated into the predictive model itself but could be a completely independent use-case of ML to ensure validity of the incoming data and quality of variables that cause dataset shifts. This would bring value no matter if the diagnostic reading is manual or computational.

5.2 Development of an automated TIL scoring system

One of the novel avenues of clinical pathology is the assessment of the tumor-immune interaction in breast cancer by scoring stromal TILs in TNBC. Addressing the increased complexity and ambiguousness in the assessments of such a biomarker is pivotal to ensure standardized care of breast cancer. Therefore, we investigated how to create a computational approach to quantify the immune-infiltration of TNBC that could overcome some of the challenges of introducing such a biomarker into clinical practice.

Before our study, existing related methods have proposed alternative metrics [99] or used other stains than H&E [81, 82] rather than adhering to the VTA clinical guideline. Concurrently with our work, two other international research groups also proposed methods for CTA [100, 101] which are consistent with VTA including validation on external cohorts. The common findings across these studies are that: (i) CTA is observed to have a good to excellent agreement with VTA, and (ii) independently associated with prognosis confirming that patients with TNBC and high CTA score have a significantly favorable survival. However, both methods are workflow 1 algorithms as they still involve the manual drawing of the tumor region before the analysis is performed compared to our workflow 3-based approach. It is yet to be seen what this means in terms of variability and implementation challenges in the clinical workflow. Moreover, there are also differences in the coarseness of the tissue-level compartmentalization, and the definition of the quantitative output variables. Our method relies heavily on segmentation models to obtain pixel-precise compartments. On the contrary, this sets higher requirements to the training labels than more rule-based approaches [101]. In line with the findings of this thesis for both dataset shifts and TIL development, both studies also agree that CTA does not solve all challenges with VTA, and there is still much research to be done in terms of handling pitfalls, further development, and clinical validation.

Even though ML-based algorithms overcome many challenges of manual assessment such as reproducibility among pathologists, it is also clear that many of the same pitfalls causing standardization issues do also affect the computational methods - both during development and deployment. This knowledge should be utilized, and therefore, we emphasize the importance of a cross-functional development team to ensure reliable computational reporting of sTIL with the end goal of progressing it into the clinic.

5.3 Addressing the bottleneck of creating training labels

During the course of this thesis, there has been immense progress to train deep neural networks without or with less training data such as self-supervised learning [67]. While these methods might also apply to pathology [102], all applications still need to be trained to solve a specific task with supervised learning. Obtaining the training labels to do so,t is probably still the single most important obstacle to develop generalizable algorithms in pathology.

Originally, we hypothesized that uncertainty methods as the once investigated for detecting dataset shifts could be used to explore uncertain data regimes, and thereby decreasing the total number of examples that should be manually annotated. The process is referred to as active learning [69]. For the TILs application, we identified key issues holding this approach back from practical use. First, the uncertainty methods did not necessarily capture all the misclassification and OOD inputs which are desirable as a proposal for training examples. Secondly, and more importantly, we noticed a high intra- and inter-observer variability when pathologists were tasked to label TILs. Therefore, no matter if active learning could propose the "right" training examples, the manual training labels would not be sufficient to create the algorithms needed to adhere to the VTA guideline. To overcome this limitation and effectively scale up the training labeling process, we showed that two IHC-based labeling schemes can obtain tissue- and cell-level labels with higher quality and consistency than manual labels.

The developed method dramatically improves the label generation process but there are still challenges associated with this method such as the vanishing differences in cellular structures between physical consecutive sections, or cell-to-cell correspondence failures due to the precision of the image registration algorithm. However, there are many potential future outlooks to streamline and improve this training label workflow.

5.4 Future research

In this thesis, we only considered computational assessment independently, and then assessed how it performs relative to pathologists. But rather than expect perfection from ML-based systems, a potentially interesting avenue of research is to assess how the combination of pathologists that use ML tools performs compared to pathologists or the algorithm alone.

Another obvious next step is to investigate the external generalizability of the TIL algorithm, and set up similar valid in- and out-of-distribution dataset shifts as in Chapter 3 to understand the pitfalls of ML-assessment of sTIL in breast cancer to an even higher degree than presented in **Appendix C**. We expect that the development dataset needs to be expanded to be multi-institutional to achieve generalization to clinical practice, but with the proposed labeling scheme, especially for the cell-level IHC, this is simply a matter of scaling the existing framework.

It is also important to recognize the other use-cases in computational pathology besides automating the reading of biomarkers and optimize aspects of the routine workflow in pathology. In particular, we are only beginning to understand the importance of TILs for cancer treatment and prognosis, and computational methods open up for further studies of the significance of the intratumoral spatial distribution of TILs, which is of significance [99]. This might allow us to understand the heterogeneity of TILs and the immune-tumor interaction in more details.

It is clear from the results presented in this thesis that uncertainty estimation in deep neural networks is not a solved issue, yet, and fundamental progress on the methods needs to happen before it can be guaranteed that pathology applications will alarm if significant dataset shifts occur not caught by other control mechanisms in the laboratory. One interesting initiative that might be the driver for novel methods is the NeurIPS *Shifts Challenge* on *Robustness and Uncertainty under Real-World Distributional Shift* that raises awareness of this important but unsolved topic.

Chapter 6

Conclusion

This thesis presents different aspects on key obstacles of bringing deep learning-based algorithms into real-world settings in pathology. Our contribution is the following: (i) Thorough investigation of several state-of-the-art methods' ability to quantify uncertainty for real-world dataset shift in pathology. (ii) Development of a fully automated algorithm for tumor infiltrating lymphocyte (TIL) assessment that adheres to all steps of the manual clinical guideline. (iii) demonstrated the effectiveness of using immunohistochemistry (IHC) to obtain both tissue- and cell-level training labels for this algorithm. (iv) reviewed the pitfalls of using ML for TIL assessment while documenting development considerations to aid future studies in this topic.

By collecting one of the largest real-life datasets in pathology in terms of studying realistic changes in data distributions, we found that current uncertainty methods could only provide reliable uncertainty estimates if used within the indication and organ included in the training set, but failed silently under any novel abnormalities. Hence, one should not expect current methods to alarm any rare incidental findings if these are not included in the training distributions, or mitigated by other systems in the laboratory.

We demonstrated that it is possible to create a fully automated H&E-based computational TIL assessment (CTA) algorithm that follows all complex aspects of a manual pathology guideline where appropriate. This was strongly enabled by having a hierarchical-based approach of both tissue- and cell-level models. To investigate the potential of the algorithm, we showed, in a large retrospective cohort, that our stromal TIL density score had both high concordance with manual scoring, and association with the prognosis of patients with triple-negative breast cancer (TNBC).

To overcome the training label bottleneck, we presented an effective way to label invasive and non-invasive epithelium as two distinct classes, and the first to show that a cell-level TIL model can be trained with labels transferred from with IHC. We showed that the quality and consistency of the labels were higher than manual labels, and one pathologist was only needed to review the labels, decreasing the time and effort needed by pathologists.

Finally, our developed algorithms do not solve all problems as they have both pitfalls of their own, and others shared with manual assessment. The pitfalls include both general pathology, methodological, and data challenges, and we provided extensive considerations for mitigating these in future development efforts.

In conclusion, we have contributed to the field of computational pathology by proposing a novel algorithm that takes into account the complexity of a real-world setting. We have shown that it is possible to automatically and quantitatively score TILs, a difficult biomarker to score manually, for an aggressive and difficult-to-treat breast cancer type. We reached this objective by breaking down clinical guidelines into multiple hierarchical and interpretable deep learning-based models. Even though the current algorithm has not been validated yet, we have shown a path to the clinic by tackling the training label constraints, shedding light on several deployment aspects, and proposed solutions to remaining obstacles. Only the future will tell to what extent models with interpretable steps like ours, end-to-end models, or a new kind of computational pathology will empower the pathologists to deliver better and more standardized patient care.

But no matter what, the future looks promising.

Appendix A

Can you trust predictive uncertainty under real dataset shifts in digital pathology?

In this appendix, we include:

Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J. D., & Dahl, A. B. (2020). Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *Proceedings of 23rd International Conference on Medical Image Computing and Computer Assisted Intervention* (pp. 824-83)

Can You Trust Predictive Uncertainty Under Real Dataset Shifts in Digital Pathology?

Jeppe Thagaard[1,2], Søren Hauberg[1], Bert van der Vegt[3], Thomas Ebstrup[2], Johan D. Hansen[2], and Anders B. Dahl[1]

[1] Technical University of Denmark, Lyngby, Denmark
[2] Visiopharm A/S, Hørsholm, Denmark
jept@dtu.dk, jth@visiopharm.com
[3] University Medical Center Groningen, Groningen, The Netherlands

Abstract. Deep learning-based algorithms have shown great promise for assisting pathologists in detecting lymph node metastases when evaluated based on their predictive accuracy. However, for clinical adoption, we need to know what happens when the test set dramatically changes from the training distribution. In such settings, we should estimate the uncertainty of the predictions, so we know when to trust the model (and when not to). Here, we i) investigate current popular methods for improving the calibration of predictive uncertainty, and ii) compare the performance and calibration of the methods under clinically relevant in-distribution dataset shifts. Furthermore, we iii) evaluate their performance on the task of out-of-distribution detection of a different histological cancer type not seen during training. Of the investigated methods, we show that deep ensembles are more robust in respect of both performance and calibration for in-distribution dataset shifts and allows us to better detect incorrect predictions. Our results also demonstrate that current methods for uncertainty quantification are not necessarily able to detect all dataset shifts, and we emphasize the importance of monitoring and controlling the input distribution when deploying deep learning for digital pathology.

Keywords: Deep learning · Digital pathology · Predictive uncertainty

1 Introduction

Motivated by the predictive performance of deep learning (DL) in research [3,21] and grand challenges [2], clinical-grade DL-tools for assisting pathologists in detection of lymph node metastases are now being developed. In clinical settings where algorithms can potentially affect medical decisions, it is crucial to know how well-calibrated the underlying model is, such that the model gives a reliable estimate of the quality of the predictions. However, there exists only limited research [4,20,22] on how different distributional shifts in pathology affect

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12261, pp. 824–833, 2020.
https://doi.org/10.1007/978-3-030-59710-8_80

the accuracy of DL-based algorithms, and these do not consider predictive uncertainty. Dataset shifts are especially relevant in pathology as pre-analytical steps can introduce large variability, and the spectrum of the target indication of an algorithm can also be broad. This makes it difficult to include the whole spectrum within the training set. Rare incidental findings, which are clinically relevant, may also be missed by an algorithm because they are outside the distribution of the training set (Fig. 1).

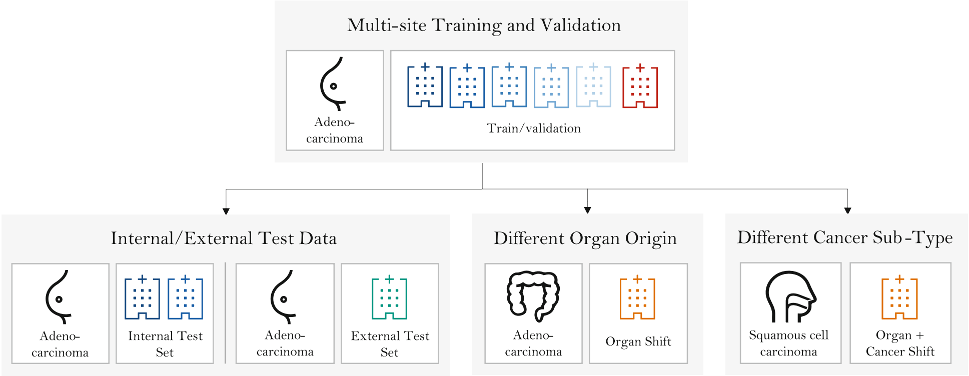


Fig. 1. Overview of experimental setup. Slides from 6 different sites are used as development data ($\mathcal{D}_{train}$ and $\mathcal{D}_{val}$), where blue (5 sites) represents CAM16-train and CAM17-train and red (one site) is DATASET2. CAM16-test defines the internal test set ($\mathcal{D}_{test,int.}$) as the 2 sites are also used as development data. DATASET3 (green) is denoted as the external test set ($\mathcal{D}_{test,ext.}$) as this site is not included in the development data. Slides from DATASET4 and DATASET5 (orange) with colon adenocarcinoma ($\mathcal{D}_{colon}$) and head and neck squamous cell carcinoma ($\mathcal{D}_{SCC}$) are used to test on different organ origin and different cancer sub-type than the original target task of detecting adenocarcinoma from breast cancer.

Our contribution is a thorough investigation of several state-of-the-art methods' ability to quantify uncertainty while keeping high accuracy. We focus on the problem of detecting cancerous tissue in digital pathology, specifically for the task of detecting lymph node metastases. This has not been covered in previous investigations such as [9,17], because the appearance and variation resulting from distributional shifts of histopathology images is very different from that of natural images. Therefore, we i) extend our evaluation to a unique real-world pathology setting with a multi-hospital single indication training set and perform an extensive evaluation on both internal and external test sets and clinically plausible distributional shifts. We ii) compare the methods in terms of performance and calibration in addition to iii) how accurate their predictive uncertainty can detect both incorrect predictions and out-of-distribution (OOD) inputs.

1.1 Related Work

Multiple popular methods have been proposed for quantifying predictive uncertainty for better calibration and robustness under distributional shifts and OOD inputs in deep neural networks (DNNs). Deep ensemble [13] is arguably the simplest method where multiple networks are trained individually and their predictions are averaged during inference. Monte Carlo Dropout (MC-Dropout) [6] is an approximate Bayesian method that uses dropout [19] during multiple forward passes during inference. Temperature scaling [7] is different as it serves as a post-processing method that learns a scaling parameter on a validation set but its performance has shown to be limited under distributional shifts [17]. Mixup [25] combines random pairs of images and their labels during training, originally aimed at increased performance but it has recently shown to improve the calibration of DNNs [23]. All methods have their advantages and limitations with regard to their complexity during training or inference.

Table 1. Details on data. * and ** denote adenocarcinoma and SCC, respectively † [14], ‡ [3].

Dataset	Purpose	No. of slides	Site
Cam16-train	Development ($\mathcal{D}_{train}$, $\mathcal{D}_{val}$)	270 (160 normal, 110 tumor*)	2 hospitals†
Cam16-test	Evaluation ($\mathcal{D}_{test,int.}$)	129 (80 normal, 49 tumor*)	2 hospitals†
Cam17-train	Development ($\mathcal{D}_{train}$, $\mathcal{D}_{val}$)	46 (0 normal, 46 tumor*)	5 hospitals‡
Dataset2	Development ($\mathcal{D}_{train}$, $\mathcal{D}_{val}$)	56 (41 normal, 15 tumor*)	Hospital-A
Dataset3	Evaluation ($\mathcal{D}_{test,\,ext.}$)	135 (67 normal, 68 tumor*)	Hospital-B
Dataset4	Evaluation ($\mathcal{D}_{colon}$)	81 (43 normal, 38 tumor*)	Hospital-C
Dataset5	Evaluation ($\mathcal{D}_{SCC}$)	60 (40 normal, 20 tumor**)	Hospital-C

2 Methods

2.1 Experimental Setup

To study a relevant application in pathology, we define the primary target task as detection of adenocarcinoma in hematoxylin and eosin (H&E) lymph node sections from breast cancer. To enable the development, we obtain datasets from public [2,3,14] and non-public sources (see details in Table 1) and evaluate both predictive accuracy and uncertainty using relevant metrics (see below).

In-distribution Shift. To evaluate whether we can trust the predictions on images not derived from the hospitals used in the development, we use Dataset3 as an external test set ($\mathcal{D}_{test,ext.}$) and Cam16$_{test}$ internal test set ($\mathcal{D}_{test,int.}$). The methods are evaluated based on their ability to generalize in terms of predictive accuracy and uncertainty.

As the same cancer sub-type can originate from different organs and metastasize to lymph nodes regardless of origin, we investigate the methods' ability to generalize to other organs than included in the training set. To enable this, we collect lymph node sections with adenocarcinoma from colon cancer ($\mathcal{D}_{colon}$).

Misclassification Detection. The ability to indicate incorrect classifications is attractive from a clinical automation perspective, so pathologists can better interfere and assess results when needed, especially when the input distribution change from the intended indication. It is easy to formulate as a binary classification problem using only the uncertainty as the prediction score, hence it is a popular downstream task to evaluate predictive uncertainty [10]. We hypothesize that current methods are better at detecting incorrect predictions when the dataset is more similar to the training distribution. To test the hypothesis, we use $\mathcal{D}_{test,int.}$, $\mathcal{D}_{test,ext.}$ and $\mathcal{D}_{colon}$ to assess the performance of the binary classification (correct vs. incorrect) on each dataset.

Out-Out-Distribution Shift. When pathologists assess lymph node sections for metastases, they are also aware of other clinically relevant abnormalities than the primary task. To mimic this setting, we collect slides that contain another histology sub-type (squamous cell carcinoma (SCC)) from head and neck cancer ($\mathcal{D}_{SCC}$), which includes both well- and un-differentiated SCCs. Since SCCs, especially well-differentiated cases, are morphological different than adenocarcinoma, we consider $\mathcal{D}_{SCC}$ a realistic out-of-distribution dataset because it contains unseen abnormalities from the same domain as the training set.

Here, our evaluation is two-fold: generalization to another cancer sub-type and the ability to detect novel classes using its predictive uncertainty. To achieve the latter, we denote all tumor regions from $\mathcal{D}_{SCC}$ as $\mathcal{D}_{out}$ and the in-distribution $\mathcal{D}_{test,ext.}$ as $\mathcal{D}_{in}$. We then compare each method to discriminate between $\mathcal{D}_{out}$ and $\mathcal{D}_{in}$.

Since poorly differentiated SCC can look morphologically similar to adenocarcinoma, we also take a subset of $\mathcal{D}_{SCC}$ diagnosed as well-differentiated SCC ($N = 5$) and treat only samples from these as OOD inputs in a final experiment.

Reference Standard. Similar to the Camelyon dataset, all ground truth annotations on the non-public datasets were carefully prepared under the supervision of expert pathologists with additional slides stained with cytokeratin immunohistochemistry (IHC). All work related to the non-public datasets was approved by their institutional review board.

2.2 Evaluation Metrics

We employ *Accuracy*, *Area Under the Receiver Operating Characteristics curve* (AUROC) and *Precision-Recall curve* (AUPR) to report classification performance (normal vs. tumor). As suggested by Guo et al. [7], we use the *Expected*

Calibration Error ECE [16] to measure the calibration for each model. First, we compute the confidence of each of N observation denoted $p(\hat{y}_n)$, and bin these into H bins. We then calculate the ECE by comparing the content of each bin to its average accuracy. Let B_h be the set of indices for bin h. We calculate the bin accuracy

$$\mathrm{acc}(B_h) = |B_h|^{-1} \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \tag{1}$$

and the bin confidence

$$\mathrm{conf}(B_h) = |B_h|^{-1} \sum_{n \in B_h} p_n(\hat{y}). \tag{2}$$

Then we get

$$\begin{aligned} \mathrm{ECE} &= \frac{1}{N} \sum_{h=1}^{H} |B_h| \cdot |\mathrm{acc}(B_h) - \mathrm{conf}(B_h)| && (3) \\ &= \frac{1}{N} \sum_{h=1}^{H} \left| \sum_{n \in B_h} p_n(y) - \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \right| && (4) \end{aligned}$$

where $\delta(x) = 1$ if $x = 0$ or $\delta(x) = 0$ if $x \neq 0$, and y_n^* is the true label.

For misclassification and OOD detection, we use also AUROC and AUPR but on the classification performance of correct vs. incorrect and in- vs. out-of-distribution, respectively. We use *False Positive Rate at 95% True Positive Rate* (FPR95) to compare method at a certain operating point. As noted by [1], these metrics are more reliable to compare for OOD detection as the task remains the same regardless of method.

2.3 Overview of Methods

We focus on methods that model $p(y|x)$ as these are the most popular in medical image analysis [3,15] and are known to scale well [12,13]. As a baseline, we use the softmax of a standard DNN to obtain posterior probabilities. For all methods, we obtain the prediction as $\hat{y} = \arg\max_y p(y|x, \theta)$ and the confidence as the maximum softmax probability $p(\hat{y}) = \max_y p(y|x, \theta)$.

MC-Dropout. We train using dropout [19] with rate p and apply L forward passes during inference with dropout enabled as described in Gal et al. [6].

Deep Ensemble. We train M standard DNNs independently of each other following [13] and combine the predictions as

$$p(y = k|x, \theta) = \frac{1}{M} \sum_{m=1}^{M} p_m(y = k|x, \theta_m) \tag{5}$$

Mixup. Recently proposed as a simple method by [25] for training better DNNs where two random input samples (x_i, x_j) and their corresponding labels (y_i, y_j) are combined using:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{6}$$

where $\lambda \in [0,1]$ determines the mixing ratio of the linear interpolation. λ is drawn from a symmetric Beta distribution Beta(α, α), where α controls the strength of the input interpolation and the label smoothing. We train a DNN with mixup using standard cross-entropy calculated on the soft-labels instead of the hard labels. We refer to [25] for the full details on mixup.

Table 2. Evaluation of predictive performance. $*\alpha = 0.3$

	$\mathcal{D}_{test,int.}$			$\mathcal{D}_{test,ext.}$			$\mathcal{D}_{colon}$		
	Acc	AUROC	AUPR	Acc	AUROC	AUPR	Acc	AUROC	AUPR
Baseline	90.5	96.5	95.1	**94.3**	97.9	94.3	**79.0**	90.7	92.8
Ensemble	90.1	**97.3**	**95.9**	**94.3**	**98.1**	**96.8**	78.1	**92.3**	**94.2**
MC-Dropout	**91.0**	97.0	95.7	93.8	97.7	96.2	78.0	90.9	93.4
Mixup*	86.5	95.6	94.2	93.4	97.1	94.6	75.8	91.0	92.6

2.4 Implementation and Training Details

We perform a train/validation split on the development dataset and use these to train and select hyper-parameters for all methods. All datasets are sampled in patches (512 × 512 pixels) at 20× magnification with 50% (strided) and 150% (overlapping) sampling fraction for normal and tumor, respectively. We employ a ResNet-50 [8] architecture as the backbone for all methods because there are negligible changes between different image classifiers [9]. We use $M = 5$ to create the ensemble as reported by [17] to be sufficient. For MC-dropout, initial experiments of different implementation variations showed no performance differences. Hence, we add a dropout before the logit layer similar to [12] with $p = 0.5$ and use $L = 50$. All models are trained for 15 epochs with ADAM [11] $(\beta = (0.9, 0.999))$ with weight decay (0.0005) using a mini-batch size of 16. We use an initial learning rate of 0.01 and drop it with factor 10 every 5th epoch for all methods except mixup which required a lower initial learning rate of 0.001 to converge. For mixup, we experimented with $\alpha \in [0.1, 0.3, 0.5, 1.0]$ and we report results with $\alpha = 0.3$ as this performed best on $\mathcal{D}_{val}$. In all experiments, we apply data augmentation similar to [15] and use Pytorch [18] and Pytorch-Lightning [5].

3 Results

3.1 Evaluating Predictive Performance Under Dataset Shifts

First, we evaluate the predictive performance on the primary task of detecting adenocarcinoma in lymph node sections. We summarize the results in Table 2, and the ROC-curves for all methods and dataset shifts are shown in Fig. 2. The results show that all methods can archive high predictive performance on both the internal and external test sets. All methods perform significantly worse when evaluated on the colon dataset $\mathcal{D}_{colon}$ with mixup performing worst. Interestingly, all methods have higher AUROC on $\mathcal{D}_{SCC}$ (see Table 4) compared to $\mathcal{D}_{colon}$ even though the cancer sub-type is histological different, especially in the well-differentiated cases. In general, deep ensemble slightly outperforms all other methods on threshold independent metrics like AUROC and AUPR.

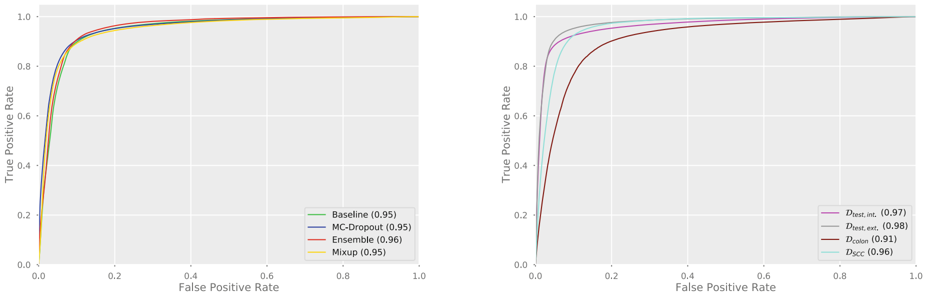


Fig. 2. ROC-curves for predictive performance. Left shows each methods with ROC curves averaged across all datasets. Right shows each dataset with ROC curves averaged across all methods.

3.2 Evaluating Predictive Uncertainty Under Dataset Shifts

We present results of calibration and detection of incorrect classified examples together in Table 3. In terms of ECE, deep ensemble and mixup improve calibration compared to the baseline method, whereas MC-dropout performs worse for the external and colon dataset. When using each method's predictive uncertainty to detect misclassifications on the test set, deep ensemble and MC-dropout have higher AUROC and AUPR on all three datasets than baseline and mixup. However, the quality of the predictive uncertainty for decreases slightly when dataset shift increases.

Table 3. Evaluation of calibration and misclassification detection. $^*\alpha = 0.3$

	$\mathcal{D}_{test,int.}$			$\mathcal{D}_{test,ext.}$			$\mathcal{D}_{colon}$		
	ECE	AUROC	AUPR	ECE	AUROC	AUPR	ECE	AUROC	AUPR
Baseline	4.9	82.6	35.7	2.1	77.7	28.6	11.8	76.7	42.0
Ensemble	**2.1**	83.9	35.6	**0.6**	**82.3**	**30.2**	**7.5**	**78.6**	**44.5**
MC-Dropout	4.6	**84.0**	35.3	2.6	79.8	29.7	13.3	77.2	43.5
Mixup*	4.2	79.1	**36.5**	0.9	80.9	29.3	9.7	71.5	41.4

3.3 Evaluating on Different Cancer Sub-type

The left part of Table 4 shows the performance on $\mathcal{D}_{SCC}$, while the right side summarizes the result of the OOD experiment. All methods show strong predictive accuracy, but fail to recognize SCC as an unseen class. Here, both ensemble and mixup outperform the baseline and MC-dropout methods.

Table 4. Evaluation of performance and OOD detection on $\mathcal{D}_{SCC}$. $^*\alpha = 0.3$

	Performance			OOD			ODD (only well-diff.)		
	Acc	AUROC	AUPR	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
Baseline	89.3	95.4	88.4	64.1	37.3	97.6	70.6	5.2	90.9
Ensemble	**89.7**	**96.3**	**91.8**	73.2	46.2	92.6	81.6	7.4	71.1
MC-Dropout	89.0	95.9	91.5	59.8	35.6	99.3	67.5	4.7	84.8
Mixup*	87.5	95.8	89.2	**86.3**	**53.6**	**47.5**	**86.5**	**8.1**	**44.6**

4 Discussion and Conclusion

We have evaluated current popular methods for predictive uncertainty on clinically relevant dataset shifts for the detection of lymph node metastases in pathology slides. All methods can generalize predictive accuracy from the internal test set to the external dataset while maintaining the quality of the predictive uncertainty. When applied to another organ, all investigated methods show both decreased performance and increased overconfidence. We have shown similar behavior when evaluated on the different cancer sub-type even-though the performance decrease was smaller than under organ shift.

As site-specific variations such as sectioning, staining and scanning variability are present in the experimental internal and external setup, we have shown that current methods are able to generalize to these sources of variability. We leave it to future work to quantify how site-specific pre-analytical variations affect the current methods as it requires a more controlled data acquisition scheme.

Our experiments show minimal benefits of MC-Dropout compared to the baseline method, and it can hurt the calibration performance on all dataset shifts. We contribute this to MC-Dropout being a too weak ensemble to achieve the same effect as a true ensemble. In general, deep ensemble increases predictive performance but also shows robustness in calibration under distributional shifts. It also displays decent capability in detecting incorrect predictions, but none of the methods are sufficient on this task. Based on the results and its simplicity, deep ensemble is an attractive method for predictive uncertainty but it comes with a computational overhead during both training and inference. Here, mixup might seem to be a cheaper alternative as our results show better calibration than baseline and MC-Dropout with a slight decrease in performance. We leave it to future work to investigate effects of different implementation of MC-Dropout and mixup extensions such as [24].

The ODD experiments indicate that adenocarcinoma and SCC, especially moderate and undifferentiated, are too similar in their morphological patterns to be treated as OOD. However, when we only assume well-differentiated SCC as an unseen class, ensemble and mixup are better to indicate the dataset shift without being sufficient for ODD detection.

Based on our results, we recommend that deep learning-based algorithms are ready for clinical implementation with reliable uncertainty estimates if used within the indication and organ included in the training set, but one should not expect current methods to alarm novel abnormalities.

Acknowledgement. The work was mainly supported by Innovation Fund Denmark (8053-00008B). Furthermore, it was partly supported by a research grant (15334) from VILLUM FONDEN, by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360) and by The Center for Quantification of Imaging Data from MAX IV (QIM) funded by The Capital Region of Denmark.

References

1. Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. arXiv preprint arXiv:2002.06470 (2020)
2. Bandi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (2019). https://doi.org/10.1109/TMI.2018.2867350
3. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA - J. Am. Med. Assoc. **318**(22), 2199–2210 (2017). https://doi.org/10.1001/jama.2017.14585
4. Ciompi, F., et al.: The importance of stain normalization in colorectal tissue classification with convolutional networks. In: ISBI, pp. 160–163 (2017)
5. Falcon, W.: Pytorch lightning. GitHub. Note. https://github.com/PyTorchLightning/pytorch-lightning (2019)
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML, pp. 1321–1330 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
10. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
11. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR, pp. 1–15 (2014)
12. Kirsch, A., van Amersfoort, J., Gal, Y.: BatchBALD: efficient and diverse batch acquisition for deep Bayesian active learning. In: NeurIPS, pp. 7026–7037 (2019)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, pp. 6402–6413 (2017)

14. Litjens, G., et al.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience **7**(6), giy065 (2018). https://doi.org/10.1093/gigascience/giy065
15. Liu, Y., et al.: Artificial intelligence based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch. Pathol. Lab. Med. **143**(7), 859–868 (2018)
16. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: AAAI (2015)
17. Ovadia, Y., et al.: Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: NeurIPS, pp. 13991–14002 (2019)
18. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035 (2019)
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
20. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. arXiv preprint arXiv:1909.11575 (2019)
21. Steiner, D.F., et al.: Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am. J. Surg. Pathol. **42**(12), 1636 (2018)
22. Tellez, D., et al.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. **58**, 101544 (2019)
23. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: improved calibration and predictive uncertainty for deep neural networks. In: NeurIPS, pp. 13888–13899 (2019)
24. Verma, V., et al.: Manifold mixup: Better representations by interpolating hidden states. In: ICLR, pp. 6438–6447 (2019)
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: ICLR (2018)

Appendix B

Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers

In this appendix, we include:

Thagaard, J., Stovgaard, E.S., Vognsen, L.G., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Vincentz, R.E., Jepsen, R.K., Roslind, A., Kümler, I., Nielsen, D., & Balslev, E. (2021) Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers. In *Cancers* 13(12):3050.

cancers

MDPI

Article

Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers

Jeppe Thagaard [1,2,*,†], Elisabeth Specht Stovgaard [3,†], Line Grove Vognsen [1,2], Søren Hauberg [1], Anders Dahl [1], Thomas Ebstrup [2], Johan Doré [2], Rikke Egede Vincentz [3], Rikke Karlin Jepsen [3], Anne Roslind [3], Iben Kümler [4], Dorte Nielsen [4] and Eva Balslev [3]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; lgv@visiopharm.com (L.G.V.); sohau@dtu.dk (S.H.); abda@dtu.dk (A.D.)

[2] Visiopharm A/S, 2970 Hørsholm, Denmark; teb@visiopharm.com (T.E.); jdh@visiopharm.com (J.D.)

[3] Department of Pathology, Herlev and Gentofte Hospital, 2730 Herlev, Denmark; elisabeth.ida.specht.stovgaard@regionh.dk (E.S.S.); rikke.egede.vincentz.02@regionh.dk (R.E.V.); rikke.karlin.jepsen@regionh.dk (R.K.J.); anne.roslind@regionh.dk (A.R.); Eva.Balslev@regionh.dk (E.B.)

[4] Department of Oncology, Herlev and Gentofte Hospital, 2730 Herlev, Denmark; Iben.Kumler@regionh.dk (I.K.); dorte.nielsen.01@regionh.dk (D.N.)

* Correspondence: jth@visiopharm.com

† Equal contributors.

Simple Summary: Around 15% of breast cancer patients are diagnosed as triple-negative (TNBC), which have significantly lower 5-year survival rates (77%) than other types of breast cancer (93%). Our study aimed at developing an image analysis-based biomarker to assess how the immune system interacts with the tumor and investigate the potential added value of stromal tumor-infiltrating lymphocytes (sTIL) for the prognosis of overall survival compared to the manual approach. In a large retrospective cohort of 257 patients, we found that our fully automated hematoxylin and eosin (H&E) image analysis pipeline can quantify sTIL density showing both high concordance with manual scoring and association with the prognosis of patients with TNBC. It also overcomes natural limitations of manual assessment that hinder clinical adoption of the immune biomarker. We conclude that sTIL scoring by automated image analysis has prognostic potential comparable to manual scoring and should be further investigated for future use in a clinical setting.

Abstract: Triple-negative breast cancer (TNBC) is an aggressive and difficult-to-treat cancer type that represents approximately 15% of all breast cancers. Recently, stromal tumor-infiltrating lymphocytes (sTIL) resurfaced as a strong prognostic biomarker for overall survival (OS) for TNBC patients. Manual assessment has innate limitations that hinder clinical adoption, and the International Immuno-Oncology Biomarker Working Group (TIL-WG) has therefore envisioned that computational assessment of sTIL could overcome these limitations and recommended that any algorithm should follow the manual guidelines where appropriate. However, no existing studies capture all the concepts of the guideline or have shown the same prognostic evidence as manual assessment. In this study, we present a fully automated digital image analysis pipeline and demonstrate that our hematoxylin and eosin (H&E)-based pipeline can provide a quantitative and interpretable score that correlates with the manual pathologist-derived sTIL status, and importantly, can stratify a retrospective cohort into two significant distinct prognostic groups. We found our score to be prognostic for OS (HR: 0.81 CI: 0.72–0.92 $p = 0.001$) independent of age, tumor size, nodal status, and tumor type in statistical modeling. While prior studies have followed fragments of the TIL-WG guideline, our approach is the first to follow all complex aspects, where appropriate, supporting the TIL-WG vision of computational assessment of sTIL in the future clinical setting.

check for updates

Citation: Thagaard, J.; Stovgaard, E.S.; Vognsen, L.G.; Hauberg, S.; Dahl, A.; Ebstrup, T.; Doré, J.; Vincentz, R.E.; Jepsen, R.K.; Roslind, A.; et al. Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers. *Cancers* **2021**, *13*, 3050. https://doi.org/10.3390/cancers13123050

Academic Editor: Katja Pinker-Domenig

Received: 12 May 2021
Accepted: 17 June 2021
Published: 18 June 2021

Keywords: deep learning; digital pathology; image analysis; prognostic biomarker; survival analysis; triple-negative breast cancer; tumor microenvironment (TME); tumor-infiltrating lymphocytes

1. Introduction

The host immune system and interactions in the tumor microenvironment (TME) play an important role in clinical outcomes for patients with triple-negative breast cancer (TNBC) [1–3]. TNBC is an aggressive and difficult-to-treat cancer type that represents approximately 15% of all breast cancers [4]. It is defined by a lack of estrogen and progesterone hormone receptors (ER/PR) and expression of human epidermal growth factor receptor 2 (HER2), i.e., common treatment options are not very effective, resulting in a lower 5-year survival rate (77%) than other types of breast cancer (93%) [5,6].

Recently, stromal tumor-infiltrating lymphocytes (sTIL) have resurfaced as a strong prognostic biomarker for overall survival (OS) [7–10], and guidelines for manual assessment have been proposed [11] to standardize reporting, increase reproducibility, and improve clinical adoption [12,13]. Nevertheless, the manual assessment has innate limitations [14] that hinder clinical adoption. These include human limitations such as inter-reader variability, bias, and limits of the routine diagnostic laboratory such as time and staff constraints, especially in remote and under-resourced settings. The International Immuno-Oncology Biomarker Working Group (TIL-WG) has therefore envisioned that computational assessment of sTIL could overcome the limitations of manual assessment and recommended that any algorithm should follow the manual guidelines where appropriate [15]. However, to the best of our knowledge, no published computational approach exists that follows all the key steps of the TIL-scoring guideline.

sTIL consists of a pool of immune cell types found in the TME such as cytotoxic CD8+ T-cells, natural killer (NK) cells, macrophages, T-helper cells, and immune-suppressing B-cells and regulatory CD4+ T-cells [16,17]. T-cells make up the majority of TILs in breast cancer [18]. It has a long history as a prognostic biomarker (more than 100 years) [19], but its clinical validity for early-stage TNBC was only recently well-established through level 1b evidence [20–22]. Incorporating sTILs into standard clinical practice is now endorsed by multiple international clinical standards since 2019 (St. Gallen Breast Cancer Expert Committee [12], World Health Organization (WHO) [23], and ESMO [24]). The guidelines to manually score sTIL status is proposed by the TIL-WG, and briefly, scored as the area of tumor-associated stroma occupied by TILs estimated as a percentage of total tumor-associated stromal area, where areas of necrosis, ductal and lobular carcinoma in situ (DCIS/LCIS), and normal breast tissue are excluded [25].

Most studies of computational TILs have employed patch- or object detection-based approaches [26–29] with manual region outlining as part of the pipeline [30]. Some of these also used multiplexed immunofluorescence (mIF) [31] or immunohistochemistry (IHC) [32,33] to classify cells as lymphocytes. All existing studies proposing H&E-based algorithms rely on only manual H&E ground truth annotations to train their model even though the manual human limitations have shown inconsistencies in this task [14]. None of these studies capture all the concepts of stromal and intratumoral TILs and account for confounding morphologies specific to different tumor sites, subtypes, and histologic patterns as envisioned by the TIL-WG [15]. Another unanswered question is the objective of an automated approach, i.e., whether the performance should be measured as the concordance between manual and automated sTIL status, the clinical outcome of the patient, or a mix of both [34].

In this study, we present a fully automated digital image analysis pipeline that integrates key aspects of the manual guideline to compute a prognostic biomarker for TNBC patients. Our approach combines both cell- and tissue-level information from whole slide images (WSIs) in both creation of ground truth annotations and during inference, which enables a robust approach that can be employed on routine H&E-stained slides. We show the existence of human inter-observer variability in the ground truth generation, and we propose to use combinatory IHC to generate more objective ground truth for both cell- and tissue-level models. We demonstrate that our H&E-based pipeline can provide a quantitative and interpretable score that correlates with the manual pathologist-derived sTIL status, and importantly, has the potential to show the prognostic implications of the sTIL status in a retrospective cohort of TNBC patients in a manner comparable to manual scoring.

2. Materials and Methods

2.1. Data Sources and Study Population

We used a cohort of patients operated for primary TNBC at Herlev and Hillerød Hospitals, Denmark, between 1 January 2004 and 31 December 2010, and who had freshly cut and stained H&E full tumor slides available. The exclusion criteria were neoadjuvant chemotherapy, previous malignancy within the past 5 years prior to diagnosis, recurrence of previous breast cancer, bilateral/multifocal breast cancer, and tumors with only microinvasion. If previous HER2 analysis had not been performed, this was conducted at the time of inclusion in the study, and patients with HER2 overexpression were excluded. A total of 262 eligible patients had freshly cut and stained H&E-stained slides from original tumor blocks from primary surgery available for analysis (a flowchart of in- and exclusion in the study can be seen in Supplementary Figure S1). Clinical information was gathered from the patient journals and/or pathology reports. A follow-up was completed on 1 July 2019. All clinical data were stored and processed at the Pathology Department, Herlev, and Gentofte Hospital, and no third party had access to data with patient information. See Supplementary Table S1 for an overview of included patients.

Patients in the inclusion period received standard chemotherapy regimens and radiation therapy if indicated. Chemotherapy regimes varied somewhat over time, as standard chemotherapy treatment in Denmark consisted of cyclophosphamide, epirubicin, and 5-fluorouracil (5-FU) from 2004 to 2007, and epirubicin, cyclophosphamide, and docetaxel from 2007 to 2010.

The H&E-staining was performed according to a well-established protocol also used in daily diagnostics at the Department of Pathology, Herlev and Gentofte Hospital, Denmark. The 4 μm slides were sectioned from formalin-fixed, paraffin-embedded (FFPE) tumor blocks and mounted on glass slides. The tissue was then deparaffinized in Tissue Clear (SAKURA Tissue Tek) and alcohol, washed with water and stained with Mayers hematoxylin (pH 2.7) and eosin (diluted with 70% alcohol), and finally treated with 99% alcohol before cover-slipping. Staining procedures varied minimally over the inclusion period, and for the digital pipeline, only freshly sectioned and stained slides were used following the procedure outlined above.

For the model development, we used only fully anonymized H&E-stained slides of TNBC tumors from Herlev Hospital, as well as publicly available slides from the TCGA-BRCA database.

The evaluation of tumor-infiltrating lymphocytes in TNBC was approved by the Danish Ethics Committee (project number H-15015306). The material used in the study was previously obtained for clinical purposes. At the time of collection, patients were informed that the material could be used for research purposes unless they registered actively in The Danish Registry for Use of Tissue. No patients included in this study had registered there.

2.2. Fully Automated Image Analysis Pipeline Design

In order to support a fully automated image analysis, we developed multiple steps into a combined algorithm: (1) we trained a convolutional neural network (CNN) to detect the tissue from the background glass slide at 5X magnification to limit the analysis to only the relevant regions; (2) a second tissue-level CNN at 10X to segment tumor, necrosis and non-invasive epithelial (normal, pre-invasive lesions); (3) an object-based density analysis of tumor regions to estimate the macro outlining of the entire tumor, hence defining the tumor-associated stroma; (4) a third cell-level CNN at 20X to detect and classify cells as TILs (mononuclear immune cells); and finally, (5) output result and local density calculation (heatmap) to quantify and visualize extracted information from the tissue- and cell-level models. The full pipeline is shown in Figure 1. All digital image analysis steps were developed and performed with the Visiopharm platform (Visiopharm A/S, Hørsholm, Denmark).

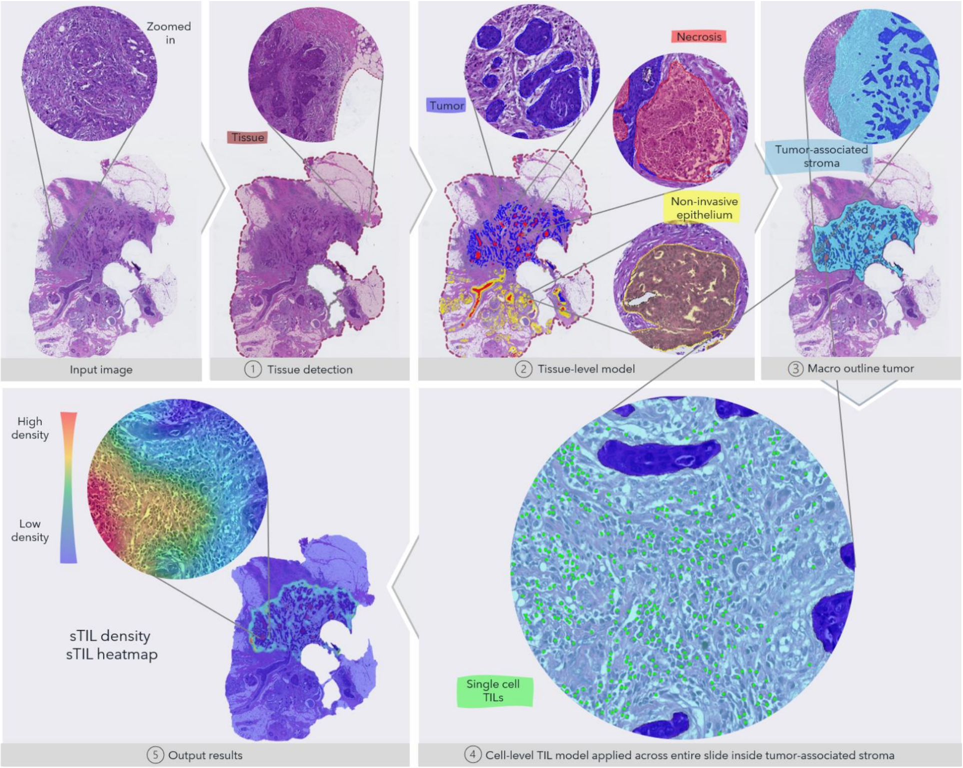


Figure 1. Overview of the fully automated image analysis pipeline. The input data are the scanned WSI of a TNBC patient, which is then analyzed by multiple steps. First, the tissue (dark red) is recognized from the glass to limit the analysis to only the relevant part of the scanned slide. Secondly, the tissue-level model classifies slide regions into tumor tissue (blue), non-invasive epithelium (yellow), and necrotic regions (red). In the third step, the macro-outline of the tumor is approximated, and then tumor-associated stroma and margin (turquoise) are defined. Cells across the entire sample in the tumor-associated stroma are classified as TILs (green) or not, and finally, the sTIL density and heatmap can be outputted for review.

We trained all CNNs with a VGG-based encoder pre-trained on ImageNet [35], where the tissue- and cell-level models use DeepLabV3 [36] and U-Net [37] inspired decoders, respectively. We applied random color augmentation (brightness, contrast, hue, and saturation), H&E stain augmentation [38], and spatial transformation (rotation, flipping). See Section 2.3 for more information on the dataset development used for these models.

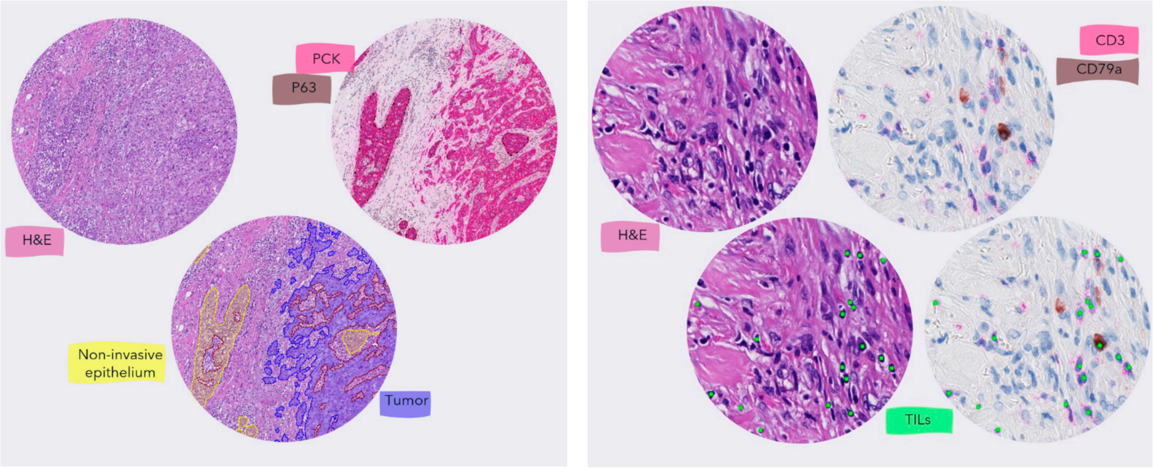
To define the tumor-associated stroma, we evaluated the local accumulated tumor area using a fixed circular kernel (radius = 750 µm) combined with morphological operations (closing/opening). The approach was designed to mimic how the pathologist would draw the macro outline of the entire tumor. We included a margin of 250 µm from the border of the tumor into the surrounding stroma. This approximation of the margin aligns with the TIL-WG guideline on including the invasive margin.

To obtain the cellular density of sTIL, we applied the cell-level TIL model across the entire macro-tumor and excluded detected TILs within regions of necrosis, a central hyalinized scar in the tumor core, tumor, and within 150 µm proximity of non-invasive epithelial to avoid dense lymphatic aggregates surrounding these regions.

Lastly, we calculated the sTIL density as the number of TILs within the tumor-associated stroma per mm^2. We also calculated the local density with a fixed circular kernel (radius = 200 µm) and visualized this as a heatmap to provide both a quantitative and visual estimate of the sTIL heterogeneity for a reviewing pathologist.

2.3. Cell and Tissue-Level Model Development

To obtain robust performance of both our tissue- and cell-level models, we developed them using an IHC-guide annotation scheme on a holdout set (n = 21 patients) from the Herlev cohort (see Figure 2) supplemented by expert pathologist annotations for the tissue-level model on a subset (n = 55 images) of the TCGA-BRCA dataset.



(**a**) Tissue-level training data (**b**) Cell-level training data

Figure 2. The process to generate objective training data. (**a**) The training annotations for the tissue-level model were generated using IHC when available. For the images from the TCGA-BRCA, the annotations were manually generated by a pathologist. (**b**) the TILs training annotations were generated as center-dot labels on cells that were either CD3 or CD79a positive to make sure that all mononuclear immune cells were included as stated by the TIL-WG guideline.

For the tissue-level model, we created new consecutive serial sections stained with H&E and pan-cytokeratin (PCK; clone AE1/AE3, DAKO Omnis) + P63 (clone DAK-P63, DAKO Omnis), respectively in the holdout set from Herlev. To generate the training data, we digitally aligned two slides using an affine registration algorithm (Tissuealign, Visiopharm A/S, Hørsholm, Denmark) and iteratively selected FOVs manually to maximize the variation in morphology of stroma, tumor, necrotic, and non-invasive regions. To increase the robustness of the model and the variation in the training data, we also included manually annotated slides from TCGA-BRCA and used the same iterative process until we saw no further performance increase on a small holdout set of the development data. We conducted the final training and validation of the tissue-level model on a ground truth dataset (n = 76 images) verified by a single pathologist (ES) before including it in the full pipeline for testing.

For the cell-level model, we only used a holdout set from Herlev as we created new sections that were first stained with H&E, then scanned, followed by removal of H&E with re-staining of a chromogenic IHC protocol (CD3 (clone F7.2.38, DAKO) and CD79a (clone JCB117, DAKO Omnis)) to highlight all mononuclear immune cells (lymphocytes and plasma cells). After digitalization, we aligned the images of the same sections as above and used a similar iterative approach to select FOVs to maximize the variation of low-, mid-, and high-density lymphocyte regions in both close and distance proximity to tumor regions. To the best of our knowledge, we are the first to apply this approach to obtain ground truth annotations for the detection and classification of TILs in H&E-stained sections. We trained and validated the final cell-level model on a ground truth dataset (n = 12 images) spanning 69 FOVs and 7277 individual lymphocytes and plasma cells. This dataset was also verified by a single pathologist (ES) reviewing all annotations with both H&E and IHC staining side-by-side.

As we deemed the cell-level model most critical to the full analysis pipeline, we conducted further testing against three expert pathologists before including it in the full pipeline, see Section 2.4 below.

2.4. Inter-Reader Variability and Validation of the Cell-Level Model

We obtained the validation set and investigated the following three key aspects; (1) the effect of having IHC available on manual recognition of a cell as a lymphocyte or not, (2) the inter-reader variability between manual readers using H&E only, and (3) the analytical performance of the cell-level TIL model. This was performed by having three pathologists mark and count sTILs. One pathologist (ES) with H&E aligned with IHC and two (RV and RJ) with H&E only to mimic the clinical setting. We used full slide images (n = 4) that were not part of the development data, where we preselected a total of 12 FOVs spanning a range of low, mid, and high-density TIL regions in intertumoral stroma varying range of proximity to tumor regions. The pathologist with access to H&E and IHC used the Visiopharm platform (Visiopharm A/S, Hørsholm, Denmark) to align the two images, so information from both could be displayed at the same time at a cellular level. The pathologists with access to only H&E used the Concentriq platform (Proscia Inc., Philadelphia, MA, USA) to mark cells as sTILs, which then could be imported to the Visiopharm platform for further analysis.

2.5. Manual Biomarker Assessment

To obtain the manual sTIL status, we used H&E slides from two FPPE tumor blocks, if available, and averaged the score or a single slide if only one block was available. Either the original H&E slides from diagnostics following primary surgery were used, or two new 4 micrometer slices were cut and stained with H&E following routine procedures. The sTIL evaluation followed guidelines published by the TIL-WG [25]. Three pathologists (ES, AR, and EB) evaluated 204 cases, and the remaining cases were evaluated by a single pathologist (ES) with a consensus reached with the other two pathologists in difficult cases. We used the manual sTIL status as a continuous variable when possible and with a cutpoint of >10% [21,39–41].

2.6. Statistical Analysis

We used overall survival (OS) as the primary endpoint for prognostic analysis, defined as the time from primary surgery until death from any cause with censoring at the last visit date. We also included relapse-free survival (RFS), defined as the time from primary surgery to local or distant relapse with censoring at death or date of the last visit, as the secondary endpoint.

We applied the Kaplan–Meier method [42] to estimate OS and RFS, and Cox proportional hazard models [43] to quantify the hazard ratio (HR) for the effects of biomarker groups (continuous or with distinct cut-offs). For continuous variables, we divided the manual sTIL with 10, and the sTIL Density with 300, so the HRs given represent differences of increments of 10 and 300, respectively.

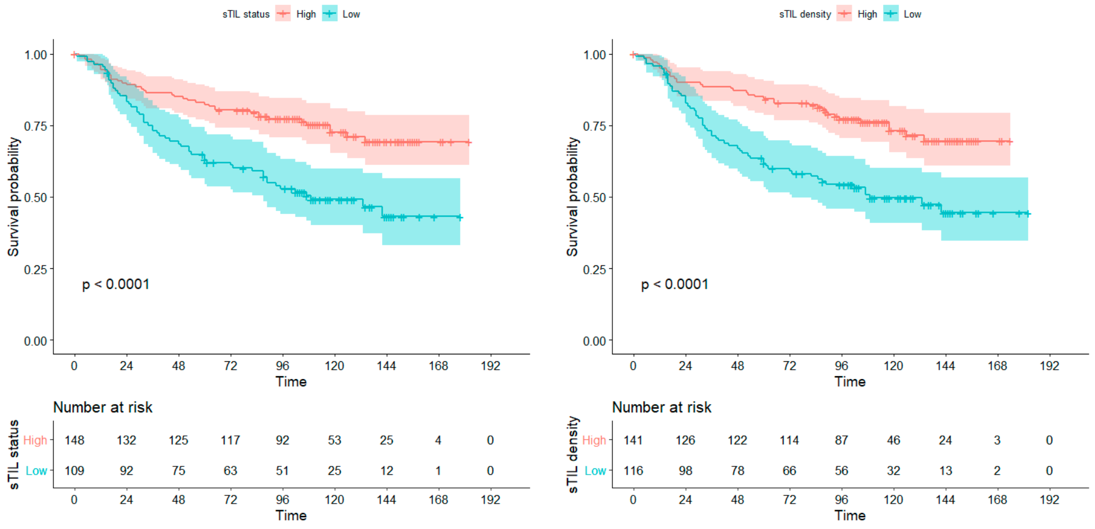
The multivariate analysis included age ($\geq$50 vs. <50 years), tumor size ($\leq$2 vs. >2 cm), number of lymph node metastases at primary surgery (0 vs. 1–3, 0 vs. $\geq$4), tumor type (ductal vs. lobular, ductal vs. other). Only cases with complete data were included in the multivariate analysis.

We conducted all statistical analyses in the R (version 4.0.3).

3. Results

3.1. Automatic sTIL Density Is Associated with Improved Overall Survival

Manually assessed sTIL is known to be associated with prognosis in TNBC patients [21,44], often stratified into two prognostic groups: high and low sTIL status [21,39,40]. To be able to investigate if the sTIL density score is similarly associated with OS, we also stratified the patient cohort into two groups: high and low sTIL density by using maximally selected rank statistics [45] for cutpoint selection of our automated approach. We found an optimal cutpoint of 470 sTIL/mm^2 and used this to estimate OS according to the Kaplan–Meier method, and compared the results to the manual sTIL status with cutpoint > 10% [21,39,40], see Figure 3. For the included cohort, both manual sTIL status and sTIL density stratified the patients significantly into two distinct prognostic groups (p < 0.0001).



(**a**) Manual assessment

(**b**) Automated image analysis

Figure 3. Overall survival estimated by Kaplan–Meier analysis. (**a**) Stratification of patients into high (red) and low (blue) group using a cutpoint of >10% on the manual sTIL status. (**b**) stratification of patients into a high (red) and low (blue) group using a cutpoint of 470 sTIL/mm^2 for the automated sTIL density.

3.1.1. Univariate Analysis

To further compare our method's association with OS, we conducted a univariate analysis on both manual sTIL status and sTIL density as a continuous variable (see Table 1). Higher sTILs scores evaluated both automatically and manually were associated with significantly prolonged OS. Every 10% or 300 sTILs/mm^2 increase in the biomarker score results in ~20% decrease in risk of death for manual (HR: 0.81 CI: 0.71–0.93) and automated score (HR 0.82 CI: 0.72–0.93), respectively. Neither of the methods was significant for RFS, with only the nodal status being significantly associated with RFS (see Table 1). Most noticeably, the univariate analysis confirmed the same significant and independent prognostic value of automated sTIL density and manual sTIL assessment as a continuous variable.

Table 1. Univariate analysis of the included clinical parameters and biomarkers. [1] Manual score is in increments of 10. [2] sTIL density is continuous but normalized to increments of 300 sTILs/mm^2.

Variable	HR (95% CI)			
	OS	*p*	RFS	*p*
Age	3.37 (1.75–6.49)	<0.001	1.83 (0.96–3.52)	0.068
Nodal status				
1–3	1.61 (1.01–2.55)	0.043	2.04 (1.16–3.57)	0.013
≥4	4.37 (2.57–7.43)	<0.001	4.33 (2.20–8.51)	<0.001
Tumor size	1.55 (1.00–2.41)	0.049	1.69 (0.98–2.93)	0.060
Tumor type				
Ductal vs. lobular	4.21 (1.32–13.44)	0.015	4.07 (0.98–16.94)	0.053
Ductal vs. other	0.95 (0.58–1.55)	0.826	0.74 (0.38–1.42)	0.367
sTIL status (manual) [1]	0.81 (0.71–0.93)	0.002	0.89 (0.77–1.02)	0.090
sTIL density (auto) [2]	0.82 (0.72–0.93)	0.002	0.87 (0.75–1.02)	0.085

3.1.2. Multivariate Analysis

To investigate the added prognostic information of sTIL density versus sTIL status to standard clinical prognostic factors, we used multivariate analysis on both OS and RFS variables (see Table 2). sTIL density was still found to be prognostic for OS (HR: 0.81 CI: 0.72–0.92 p = 0.001) independent of age, tumor size, nodal status, and tumor type. The same was observed for manual sTIL status (HR: 0.79 CI: 0.68–0.91 p = 0.001). For RFS, both methods were found to be significant.

Table 2. Multivariate analysis: [1] Manual score is in increments of 10. [2] sTIL Density is continuous but normalized to increments of 300 sTILs/mm^2.

Method	Overall Survival			Relapse Free Survival		
	HR	95% CI	*p*-Value	HR	95% CI	*p*-Value
sTIL (manual) [1]	0.79	0.68–0.91	0.001	0.84	0.71–0.99	0.037
Tumor Size	1.44	0.92–2.25	0.115	1.57	0.89–2.75	0.117
Age	2.96	1.52–5.77	0.001	1.72	0.88–3.35	0.112
Nodal status						
1–3	1.92	1.20–3.07	0.007	2.23	1.26–3.95	0.006
≥4	4.52	2.61–7.84	<0.001	4.42	2.19–8.90	<0.001
Tumor type						
Ductal vs. lobular	1.79	0.55–5.84	0.335	1.73	0.40–7.46	0.461
Ductal vs. other	0.91	0.55–1.51	0.718	0.74	0.38–1.45	0.384
sTIL density (auto) [2]	0.81	0.72–0.92	0.001	0.86	0.75–1.00	0.047
Tumor Size	1.43	0.91–2.24	0.124	1.56	0.89–2.75	0.122
Age	3.02	1.55–5.90	0.001	1.76	0.90–3.43	0.099
Nodal status						
1–3	1.91	1.19–3.07	0.007	2.22	1.25–3.92	0.006
≥4	4.12	2.40–7.08	<0.001	4.11	2.06–8.19	<0.001
Tumor type						
Ductal vs. lobular	2.15	0.66–6.95	0.203	2.00	0.47–8.52	0.347
Ductal vs. other	0.89	0.54–1.48	0.664	0.74	0.38–1.44	0.375

3.2. Cell-Level TIL Model Correlates with Manual Expert Pathologists

Previous studies have shown inter-reader variability for identifying individual sTILs in H&E [14,46]. Therefore, a key part of the fully automated pipeline is to be able to count the correct number of sTILs. To determine the degree of inter-reader variability and the analytical validation of the cell-level TIL model, we used the data described in Section 2.3, where we also applied the TIL model to the same regions to measure the agreement. The results are shown in Figure 4 of the correlation between the approaches. The TIL model had a high correlation with all three pathologists, especially the pathologist with access to both H&E and IHC CD3 + CD79a (Spearman correlation coefficient $r_s = 0.916$). Moreover, the inter-reader agreement between the pathologist was also high, but with the lowest correlation between the pathologist with access IHC and pathologist 3 ($r_s = 0.783$). The lowest correlation to the TIL model was seen between pathologist 3 ($r_s = 0.853$), where the pathologist counted fewer TILs in many cases. Overall, we observed an inter-reader variability between the expert pathologists and that the TIL model had the highest correlation with the pathologist who had access to the same information (H&E + IHC) as the TIL model was trained against.

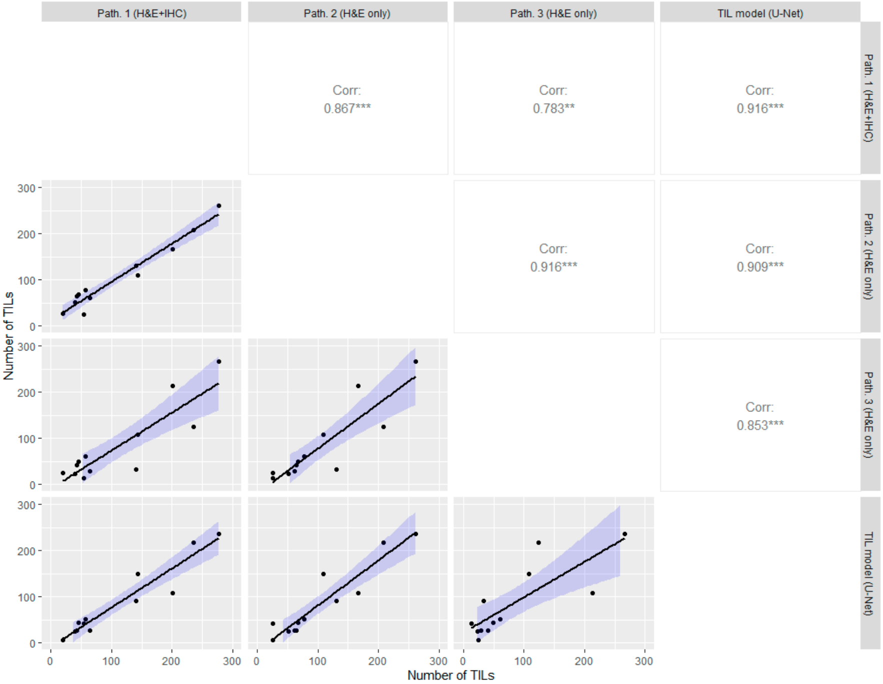


Figure 4. Inter-method variability of cell-level discrimination of TILs between the pathologist with both H&E and IHC, the two pathologists with only H&E, and our image analysis approach on a holdout test set. The lower left of the diagonal shows the correlations plot, and the upper right shows the Spearman correlation coefficient for each comparison. The asterisks ** ($p \leq 0.01$) and *** ($p \leq 0.001$) indicate the significance levels of the statistical correlation test.

3.3. Automatic sTIL Density Correlates with Manual sTIL Assessment on Full Section H&E Slides

When scaling sTIL scoring up to the full tissue section, the manual assessment score is prone to many pitfalls [14] even though guidelines are followed. To validate the full automated analysis pipeline, we used Spearman correlation to test if there is a significant linear relationship between the manual sTIL assessment score (see Section 2.5) and the automatic sTIL density output from our approach, see Figure 5. We observed a significantly high correlation ($r_s = 0.79$, $p < 0.001$) between the two methods. As expected, we did not see a perfect correlation as our method uses the computed sTILs per mm^2, whereas the manual scoring guideline is an estimate of area coverage by sTIL. We also observed larger disagreement for higher sTIL scores comparable to the inter-pathologist agreement for manual scoring whole section cases [47]. The result is comparable to the variance observed between pathologists scoring sTIL [14,47].

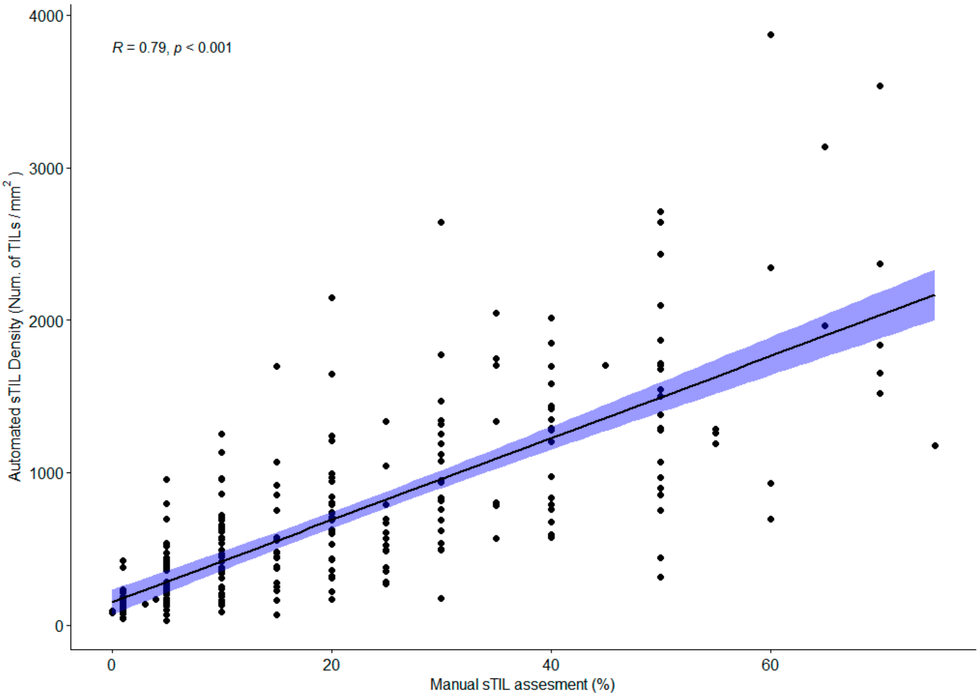


Figure 5. Correlation between manual sTIL assessment and automated sTIL density.

We found a total of 50 discrepant cases between low and high sTIL groups using the cutpoints for each method. At this specific cutpoint, this binary classification corresponds to a sensitivity and specificity of 81.2% and 80.5%, respectively (22 false positives and 28 false negatives). To understand these discrepant cases more, we looked at the manual score and image analysis quality. For 39 of the discrepant cases, the manual score was obtained as a consensus between 3 pathologists. The remaining 11 cases were scored by a single pathologist. Twenty-eight cases were scored >10% manually but are below the cutpoint for the automated method. For these, the average manual sTIL status is slightly above the cutpoint (μ = 21%) with an average standard deviation between pathologists of 5%, and the average sTIL density is 310 cells/mm^2. For the other scenario, where 22 cases were scored $\leq$10% but were above the cutpoint for the automated method, the manual sTIL status was 10% for 82% of these cases (μ = 8.6%) with an average standard deviation between pathologist of 2%. The automated sTIL density of these cases is 725 cells/mm^2. For 47 of the discrepant cases (94%), both scores from the manual and automated method were around their respective cutpoints, and we consider these within the expected discrepancy around cutpoints. The last three cases all had manual sTIL > 30% but were below the automated cutpoint. One case had a sectioning artifact resulting in a lower automated score. The two others had high lymphocyte infiltration along the invasive margin but almost no sTILs in the central tumor-associated stroma. The discrepancy might result from how the contribution from the two compartments was averaged as the automated method does not treat the two compartments (invasive margin and tumor-associated stroma) equally but averages the density across all tumor-associated stroma.

4. Discussion

In this study, we designed a digital image analysis pipeline that joins several algorithmic steps, including a tissue-level segmentation model and a cell-level TIL model that combined adhere to the manual scoring guideline by the TIL-WG. We demonstrated how our sTIL density score is independently prognostically significant for OS, similar to manual sTIL status on whole sections. Furthermore, the automatic score stratifies patients in low- and high-sTIL density groups that are highly associated with OS and correlate highly with the manual sTIL assessment. Our study shows for the first time that sTIL density in TNBC can reliably be assessed by a fully automatic deep learning pipeline.

Compared to prior attempts to apply image analysis for computational assessment of sTIL, such as patch- [26], object- [28,29], or segmentation-based methods [27,48], our study incorporates all parts of the TIL-WG guideline; from discriminating tissue from glass, and excluding necrotic regions and inflammation related to the non-invasive epithelium, such normal glands and DCIS/LCIS. A recent study [33] investigated several aspects of computational TIL assessment for prognosis in TNBC. To find the optimal compartment (margin, tumor-associated stroma, etc.), they used manual annotations and found no difference in the various regions. To investigate the immune cell population that is optimally for prognostic biomarker assessment, they used IHC for CD3, CD8, and FOXP3, and again found that all subtypes of markers correlate with survival. These observations are in line with ours as we do not discriminate between invasive margin and tumor-associated stroma but simply perform a combined assessment of the two compartments. Similarly, we do not discriminate between the immune cell subtypes but quantify all mononuclear immune cells as one class as stated by the TIL-WG guideline. These observations indicate that manual region annotations and immune cell subtypes are not necessary to obtain a prognostic immune-related biomarker for TNBC.

Recent studies have also shown the benefit of combining tissue- and cell-level deep learning models to interrogate the TME in breast cancer, such as the local TIL infiltration around DCIS structures [49], or engineering hundreds of features from these models to predict molecular signatures [50]. Our results align well with the benefits of having both multi-level analyses. In contrast to these studies, we focus on a single proven biomarker, and we sought to translate the manual guideline into a computational approach that could be performed by a computer. This can be combined with other biomarkers such as the tumor stroma ratio (TSR) [51] directly from the same H&E section, which also is associated with survival when calculated computationally on tissue microarrays (TMAs) [52], or with IHC markers such as the expression of programmed death-ligand 1 (PD-L1) [53].

To not be limited by expensive and subjective expert annotations in the development data used in this and future studies, we also rigorously focused on an objective approach to generate ground truth data that is scalable at both tissue- and cell-level. Other related applications also used similar IHC techniques to transfer annotations to H&E. Tellez et al. [38] used PHH3 to guide annotations of mitotic cells in breast cancer tissue, Bulten et al. used P63 and CK8/18 as the reference standard for a CNN to segment epithelium in prostate cancer [54], and Valkonen et al. [55] automatically transferred CD45 to an H&E slide to segment leukocytes in papillary thyroid carcinoma. Similar to ours, these methods also involve a manual step in the process. However, we use it to generate tissue- and cell-level annotations and show that this technique works for guiding annotations of all relevant mononuclear immune cells in breast cancer.

Our approach allows us to investigate and quantify the TME for a specific cellular biomarker across the entire WSI image. Hence, it overcomes the limiting constraints of manual reading as counting all cells and measuring precise stromal area in samples with complex tumor patterns is intractable to perform for a human, e.g., related to the heterogeneity in sTIL distribution [14]. Even though small differences exist in the averaging compartments between our method and the TIL-WG guideline, the sTIL density shows similar potential as a prognostic biomarker as the manual assessment for the investigated cohort. These findings also confirm previous studies in breast cancer, in which sTIL assessment is found to be associated with improved prognosis [21,44]. One of the sources for variability in manual scoring is the adherence to the guideline definition [14]. Using a computational approach that adheres to that definition increases the standardization for scoring TNBC patients, while it also shows similar concordance to the clinical outcome of those patients.

Our study also has several limitations. First, even though our models show good generalizability on the retrospective cohort (*n* = 480 WSIs), we developed them on a limited number of cases. This means that the models might not perform optimally on another study cohort from a different site with a distributional shift in, e.g., preanalytical protocols, staining protocol, or scanner type [56,57]. Future development of our approach should extend the development dataset of both tissue- and cell-level models to be multi-institutional, covering the innate variability of the above-mentioned factors.

The cutpoint for the low- and high-sTIL density also has limitations as it was found within the single study cohort. As we used the biomarker as a continuous variable in the multivariate analysis for OS, this should not affect the evidence of our methods' association to improved prognosis. The discrepancy at the binary cutpoint between the manual and automated approach should also be compared to the variability of manual scoring (intraclass correlation coefficients of 0.77–0.94 for discrete cut-off values) [14]. However, in future validation, the optimal cutpoint should be investigated further and tested on an independent cohort. In general, new emerging biomarkers must be co-developed with a digital image analysis tool to ease the clinical adoption by pathologists. By doing so, clinicians simultaneously learn about the biomarker and familiarize themselves with the pros (and cons) of quantifying it using machine learning (ML)-based scoring approaches. Hence, the clinical validation will become a combination of the biomarker and automated scoring method providing a combined computational biomarker, and not just a digital tool add-on after years of manually scoring the biomarker. With the current pace of advancement in ML for healthcare, it will also become instrumental that existing clinicians and future generations of physicians obtain formal training in computational approaches so they can better assess the clinical needs, advice on how it is best integrated into their workflow, and perform the critical appraisal of the performance of ML-based systems [58]. All this to ensure the added value in day-to-day clinical decision making.

Even though our analytical validation of the TIL model shows a high correlation between our approach and the expert pathologist, this step of the algorithm is critical to the validity of the full pipeline. There are recent efforts by regulatory instances to develop and provide the dataset for validating exactly this kind of computational step [46]. We recommend that such efforts might be supplemented by our annotation approach to generate a more objective ground truth for estimating the density of sTIL in breast cancer, so the reliance on large-scale pathologist annotation is limited while mitigating variability in the process.

Should the automated approach then completely replace the manual sTIL assessment? No. The automated approach might be faster and more reproducible in many aspects but also has several limitations, as discussed above. We recommend using our approach as another tool in the pathologist toolbox to help increase reproducibility and handle key factors such as sTIL heterogeneity by automatically computing objective counts and area metrics recognized by the models. This is also the recommendation from the TIL-WG [15]. As the diagnostic responsibility resides with the pathologist, these metrics need to be presented quantitatively and visually for manual review and sign-off. Future development of our approach could therefore extend to investigate the impact of a combined setup of a pathologist using a computational method on the clinical outcome of the patient.

5. Conclusions

We demonstrated in a large retrospective cohort that a fully automated H&E image analysis pipeline could quantify sTIL density showing both high concordance with manual scoring and association with the prognosis of patients with TNBC. While prior studies have followed fragments of the TIL-WG guideline, our approach follows all complex aspects where appropriate supporting the TIL-WG vision of computational assessment of sTIL in the future clinical setting.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/cancers13123050/s1, Figure S1: Flowchart of patients included in the study, Table S1: Clinicopathological characteristics of the patient population.

Author Contributions: Conceptualization, J.T., E.S.S., S.H., A.D., T.E., J.D., and E.B.; methodology, J.T., E.S.S., S.H., and A.D.; software, J.T., T.E. and J.D.; validation, J.T., E.S.S., E.B., R.K.J., R.E.V., and A.R.; formal analysis, J.T. and E.S.S.; data curation, L.G.V., E.S.S., R.K.J., R.E.V., E.B., A.R., I.K., and D.N.; writing—original draft preparation, J.T. and E.S.S.; writing—review and editing, all; visualization, J.T.; supervision, S.H., A.D., T.E., J.D., and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark, grant number 8053-00008B.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Danish Ethics Committee (project number H-15015306).

Informed Consent Statement: The material used in the study had previously been obtained for clinical purposes. At the time of collection, patients had been informed that the material could be used for research purposes unless they registered actively in The Danish Registry for Use of Tissue. No patients included in this study had registered there.

Data Availability Statement: Data from the Cancer Genome Atlas dataset are publicly available at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed on 19 October 2020). Restrictions apply to the Herlev dataset from Herlev and Gentofte Hospital, Denmark, and are not publicly available under the current research approval from the institutional review board and without a data processing agreement as stated by local law.

Conflicts of Interest: J.T., L.G.V., and T.E. are employees of Visiopharm, A/S. J.D. is the co-founder of Visiopharm A/S. All other authors declare no conflict of interest.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
2. Cavallo, F.; De Giovanni, C.; Nanni, P.; Forni, G.; Lollini, P.L. 2011: The immune hallmarks of cancer. *Cancer Immunol. Immunother.* **2011**, *60*, 319–326. [CrossRef]
3. Bianchini, G.; Gianni, L. The immune system and response to HER2-targeted treatment in breast cancer. *Lancet Oncol.* **2014**, *15*, e58–e68. [CrossRef]
4. Foulkes, W.D.; Smith, I.E.; Reis-Filho, J.S. Triple-Negative Breast Cancer. *N. Engl. J. Med.* **2010**, *363*, 1938–1948. [CrossRef]
5. Plevritis, S.K.; Munoz, D.; Kurian, A.W.; Stout, N.K.; Alagoz, O.; Near, A.M.; Lee, S.J.; Broek, J.J.V.D.; Huang, X.; Schechter, C.B.; et al. Association of Screening and Treatment with Breast Cancer Mortality by Molecular Subtype in US Women, 2000–2012. *JAMA* **2018**, *319*, 154–164. [CrossRef]
6. Costa, R.L.B.; Gradishar, W.J. Triple-Negative Breast Cancer: Current Practice and Future Directions. *J. Oncol. Pract.* **2017**, *13*, 301–303. [CrossRef] [PubMed]
7. Savas, P.P.; Salgado, R.; Denkert, C.; Sotiriou, C.; Darcy, P.K.P.; Smyth, M.; Loi, S. Clinical relevance of host immunity in breast cancer: From TILs to the clinic. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 228–241. [CrossRef]
8. Hammerl, D.; Smid, M.; Timmermans, A.M.; Sleijfer, S.; Martens, J.W.M.; Debets, R. Breast cancer genomics and immuno-oncological markers to guide immune therapies. *Semin. Cancer Biol.* **2018**, *52*, 178–188. [CrossRef]
9. Hudeček, J.; Voorwerk, L.; van Seijen, M.; Nederlof, I.; de Maaker, M.; Berg, J.V.D.; van de Vijver, K.K.; Sikorska, K.; Adams, S.; Demaria, S.; et al. Application of a risk-management framework for integration of stromal tumor-infiltrating lymphocytes in clinical trials. *NPJ Breast Cancer* **2020**, *6*, 1–8. [CrossRef]
10. Adams, S.; Gray, R.J.; Demaria, S.; Goldstein, L.; Perez, E.A.; Shulman, L.N.; Martino, S.; Wang, M.; Jones, V.E.; Saphner, T.J.; et al. Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. *J. Clin. Oncol.* **2014**, *32*, 2959–2966. [CrossRef] [PubMed]
11. Salgado, R.; Denkert, C.; Demaria, S.; Sirtaine, N.; Klauschen, F.; Pruneri, G.; Wienert, S.; Van den Eynden, G.; Baehner, F.L.; Penault-Llorca, F.; et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: Recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **2015**, *26*, 259–271. [CrossRef]
12. Morigi, C. Highlights of the 16th St Gallen International Breast Cancer Conference, Vienna, Austria, 20–23 March 2019: Personalised treatments for patients with early breast cancer. *Ecancermedicalscience* **2019**, *13*, 924. [CrossRef]
13. Balic, M.; Thomssen, C.; Würstlein, R.; Gnant, M.; Harbeck, N. St. Gallen/Vienna 2019: A Brief Summary of the Consensus Discussion on the Optimal Primary Breast Cancer Treatment. *Breast Care* **2019**, *14*, 103–110. [CrossRef]
14. Kos, Z.; Roblin, E.; Kim, R.S.; Michiels, S.; Gallas, B.D.; Chen, W.; van de Vijver, K.K.; Goel, S.; Adams, S.; Demaria, S.; et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer* **2020**, *6*, 1–16. [CrossRef]

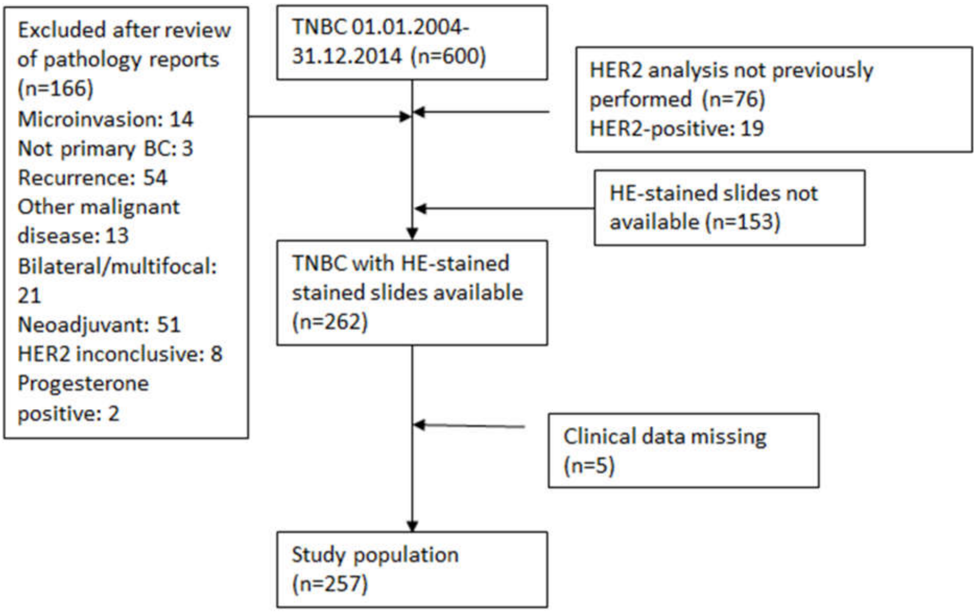
15. Amgad, M.; Stovgaard, E.S.; Balslev, E.; Thagaard, J.; Chen, W.; Dudgeon, S.; Sharma, A.; Kerner, J.K.; Denkert, C.; Yuan, Y.; et al. Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group. *NPJ Breast Cancer* **2020**, *6*, 1–13. [CrossRef] [PubMed]
16. Savas, P.; Virassamy, B.; Ye, C.; Salim, A.; Mintoff, C.P.; Caramia, F.; Salgado, R.; Byrne, D.J.; Teo, Z.L.; Dushyanthen, S.; et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **2018**, *24*, 986–993. [CrossRef]
17. Dushyanthen, S.; Beavis, P.; Savas, P.; Teo, Z.L.; Zhou, C.; Mansour, M.; Darcy, P.K.; Loi, S. Relevance of tumor-infiltrating lymphocytes in breast cancer. *BMC Med.* **2015**, *13*, 202. [CrossRef] [PubMed]
18. Ruffell, B.; Au, A.; Rugo, H.S.; Esserman, L.J.; Hwang, E.S.; Coussens, L.M. Leukocyte composition of human breast cancer. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 2796–2801. [CrossRef] [PubMed]
19. Sistrunk, W.E.; MacCarty, W.C. Life expectancy following radical amputation for carcinoma of the breast: A clinical and pathologic study of 218 cases. *Ann. Surg.* **1922**, *75*, 61–69.
20. Simon, R.M.; Paik, S.; Hayes, D.F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.* **2009**, *101*, 1446–1452. [CrossRef]
21. Loi, S.; Drubay, D.; Adams, S.; Pruneri, G.; Francis, P.A.; Lacroix-Triki, M.; Joensuu, H.; Dieci, M.V.; Badve, S.; Demaria, S.; et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. *J. Clin. Oncol.* **2019**, *37*, 559–569. [CrossRef]
22. Denkert, C.; Von Minckwitz, G.; Darb-Esfahani, S.; Lederer, B.; Heppner, B.I.; Weber, K.E.; Budczies, J.; Huober, J.; Klauschen, F.; Furlanetto, J.; et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **2018**, *19*, 40–50. [CrossRef]
23. International Agency for Research on Cancer. *WHO Classification of Tumours Series, Breast Tumours*, 5th ed.; WHO Classification of Tumours Editorial Board: Lyon, France, 2019; Volume 2. Available online: https://tumourclassification.iarc.who.int/chapters/32 (accessed on 11 April 2021).
24. Cardoso, F.; Kyriakides, S.; Ohno, S.; Penault-Llorca, F.; Poortmans, P.; Rubio, I.; Zackrisson, S.; Senkus, E. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2019**, *30*, 1194–1220. [CrossRef] [PubMed]
25. Hendry, S.; Salgado, R.; Gevaert, T.; Russell, P.A.; John, T.; Thapa, B.; Christie, M.; van de Vijver, K.; Estrada, M.V.; Gonzalez-Ericsson, P.I.; et al. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method from the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma in situ, Metastatic Tumor Deposits and Areas for Further Research. *Adv. Anat. Pathol.* **2017**, *24*, 235–251. [CrossRef] [PubMed]
26. Saltz, J.; Gupta, R.; Hou, L.; Kurc, T.; Singh, P.; Nguyen, V.; Samaras, D.; Shroyer, K.R.; Zhao, T.; Batiste, R.; et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **2018**, *23*, 181–193.e7. [CrossRef]
27. Amgad, M.; Sarkar, A.; Srinivas, C.; Redman, R.; Ratra, S.; Bechert, C.J.; Calhoun, B.C.; Mrazeck, K.; Kurkure, U.; Cooper, L.A.D.; et al. Joint Region and Nucleus Segmentation for Characterization of Tumor Infiltrating Lymphocytes in Breast Cancer. In *Medical Imaging 2019: Digital Pathology*; International Society for Optics and Photonics: Bellingham, WA, USA, 2019; p. 109560M. [CrossRef]
28. Yuan, Y.; Failmezger, H.; Rueda, O.M.; Ali, H.R.; Gräf, S.; Chin, S.-F.; Schwarz, R.F.; Curtis, C.; Dunning, M.J.; Bardwell, H.; et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **2012**, *4*, 157ra143. [CrossRef] [PubMed]
29. Basavanhally, A.N.; Ganesan, S.; Agner, S.; Monaco, J.P.; Feldman, M.D.; Tomaszewski, J.E.; Bhanot, G.; Madabhushi, A. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 642–653. [CrossRef]
30. Le, H.; Gupta, R.; Hou, L.; Abousamra, S.; Fassler, D.; Torre-Healy, L.; Moffitt, R.A.; Kurc, T.; Samaras, D.; Batiste, R.; et al. Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer. *Am. J. Pathol.* **2020**, *190*, 1491–1504. [CrossRef] [PubMed]
31. He, T.-F.; Yost, S.E.; Frankel, P.H.; Dagis, A.; Cao, Y.; Wang, R.; Rosario, A.; Tu, T.Y.; Solomon, S.; Schmolze, D.; et al. Multi-panel immunofluorescence analysis of tumor infiltrating lymphocytes in triple negative breast cancer: Evolution of tumor immune profiles and patient prognosis. *PLoS ONE* **2020**, *15*, e0229955. [CrossRef]
32. Swiderska-Chadaj, Z.; Pinckaers, H.; van Rijthoven, M.; Balkenhol, M.; Melnikova, M.; Geessink, O.; Manson, Q.; Sherman, M.; Polonia, A.; Parry, J.; et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Med. Image Anal.* **2019**, *58*, 101547. [CrossRef]
33. Balkenhol, M.C.; Ciompi, F.; Świderska-Chadaj, Ż.; van de Loo, R.; Intezar, M.; Otte-Höller, I.; Geijs, D.; Lotz, J.; Weiss, N.; de Bel, T.; et al. Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics. *Breast* **2021**, *56*, 78–87. [CrossRef] [PubMed]
34. Brown, L.C.; Salgado, R.; Luen, S.J.; Savas, P.; Loi, S. Tumor-Infiltrating Lymphocyctes in Triple-Negative Breast Cancer: Update for 2020. *Cancer J.* **2021**, *27*, 25–31. [CrossRef] [PubMed]
35. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

36. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
38. Tellez, D.; Balkenhol, M.; Otte-Holler, I.; van de Loo, R.; Vogels, R.; Bult, P.; Wauters, C.; Vreuls, W.; Mol, S.; Karssemeijer, N.; et al. Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. *IEEE Trans. Med. Imaging* **2018**, *37*, 2126–2136. [CrossRef]
39. McShane, L.M.; Altman, D.G.; Sauerbrei, W.; Taube, S.E.; Gion, M.; Clark, G.M. Reporting recommendations for tumor marker prognostic studies (REMARK). *J. Natl. Cancer Inst.* **2005**, *97*, 1180–1184. [CrossRef]
40. Fuchs, T.L.; Pearson, A.; Pickett, J.; Diakos, C.; Dewar, R.; Chan, D.; Guminski, A.; Menzies, A.; Baron-Hay, S.; Sheen, A.; et al. Why pathologists and oncologists should know about tumour-infiltrating lymphocytes (TILs) in triple-negative breast cancer: An Australian experience of 139 cases. *Pathology* **2020**, *52*, 515–521. [CrossRef] [PubMed]
41. Yuan, Y.; Lee, J.S.; Yost, S.E.; Li, S.M.; Frankel, P.H.; Ruel, C.; Schmolze, D.; Robinson, K.; Tang, A.; Martinez, N.; et al. Phase II Trial of Neoadjuvant Carboplatin and Nab-Paclitaxel in Patients with Triple-Negative Breast Cancer. *Oncologist* **2020**, *26*, e382–e393. [CrossRef] [PubMed]
42. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481. [CrossRef]
43. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **1972**, *34*, 187–220. [CrossRef]
44. Salgado, R.; Denkert, C.; Campbell, C.; Savas, P.; Nuciforo, P.; Aura, C.; de Azambuja, E.; Eidtmann, H.; Ellis, C.E.; Baselga, J.; et al. Tumor-Infiltrating Lymphocytes and Associations With Pathological Complete Response and Event-Free Survival in HER2-Positive Early-Stage Breast Cancer Treated With Lapatinib and Trastuzumab: A Secondary Analysis of the NeoALTTO Trial. *JAMA Oncol.* **2015**, *1*, 448–454. [CrossRef]
45. Lausen, B.; Schumacher, M. Maximally selected rank statistics. *Biometrics* **1992**, *48*, 73–85. [CrossRef]
46. Dudgeon, S.N.; Wen, S.; Hanna, M.G.; Gupta, R.; Amgad, M.; Sheth, M.; Marble, H.; Huang, R.; Herrmann, M.D.; Szu, C.H.; et al. A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study. preprint. *arXiv* **2020**, arXiv:2010.06995.
47. Kim, R.S.; Song, N.; Gavin, P.; Salgado, R.; Bandos, H.; Kos, Z.; Floris, G.; Eynden, G.G.G.M.V.D.; Badve, S.; Demaria, S.; et al. Stromal Tumor-infiltrating Lymphocytes in NRG Oncology/NSABP B-31 Adjuvant Trial for Early-Stage HER2-Positive Breast Cancer. *J. Natl. Cancer Inst.* **2019**, *111*, 867–871. [CrossRef] [PubMed]
48. Amgad, M.; Elfandy, H.; Hussein, H.; Atteya, L.A.; Elsebaie, M.A.T.; Elnasr, L.S.A.; Sakr, R.A.; Salem, H.S.E.; Ismail, A.F.; Saad, A.; et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **2019**, *35*, 3461–3467. [CrossRef] [PubMed]
49. Narayanan, P.L.; Raza, S.E.A.; Hall, A.H.; Marks, J.R.; King, L.; West, R.B.; Hernandez, L.; Guppy, N.; Dowsett, M.; Gusterson, B.; et al. Unmasking the immune microecology of ductal carcinoma in situ with deep learning. *NPJ Breast Cancer* **2021**, *7*, 19. [CrossRef]
50. Diao, J.A.; Wang, J.K.; Chui, W.F.; Mountain, V.; Gullapally, S.C.; Srinivasan, R.; Mitchell, R.N.; Glass, B.; Hoffman, S.; Rao, S.K.; et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.* **2021**, *12*, 1–15. [CrossRef]
51. Wu, J.; Liang, C.; Chen, M.; Su, W. Association between tumor-stroma ratio and prognosis in solid tumor patients: A systematic review and meta-analysis. *Oncotarget* **2016**, *7*, 68954–68965. [CrossRef]
52. Millar, E.K.; Browne, L.H.; Beretov, J.; Lee, K.; Lynch, J.; Swarbrick, A.; Graham, P.H. Tumour Stroma Ratio Assessment Using Digital Image Analysis Predicts Survival in Triple Negative and Luminal Breast Cancer. *Cancers* **2020**, *12*, 3749. [CrossRef]
53. Wimberly, H.; Brown, J.R.; Schalper, K.; Haack, H.; Silver, M.R.; Nixon, C.; Bossuyt, V.; Pusztai, L.; Lannin, D.R.; Rimm, D.L. PD-L1 Expression Correlates with Tumor-Infiltrating Lymphocytes and Response to Neoadjuvant Chemotherapy in Breast Cancer. *Cancer Immunol. Res.* **2015**, *3*, 326–332. [CrossRef]
54. Bulten, W.; Bándi, P.; Hoven, J.; Van De Loo, R.; Lotz, J.; Weiss, N.; Van Der Laak, J.; Van Ginneken, B.; De Kaa, C.H.-V.; Litjens, G. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **2019**, *9*, 1–10. [CrossRef]
55. Stenman, S.E.; Bychkov, D.; Kucukel, H.; Linder, N.; Haglund, C.; Arola, J.; Lundin, J. Antibody Supervised Training of a Deep Learning Based Algorithm for Leukocyte Segmentation in Papillary Thyroid Carcinoma. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 422–428. [CrossRef] [PubMed]
56. Thagaard, J.; Hauberg, S.; Van Der Vegt, B.; Ebstrup, T.; Hansen, J.D.; Dahl, A.B. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2020; pp. 824–833. [CrossRef]
57. Swiderska-Chadaj, Z.; De Bel, T.; Blanchet, L.; Baidoshvili, A.; Vossen, D.; Van Der Laak, J.; Litjens, G. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Sci. Rep.* **2020**, *10*, 1–14. [CrossRef] [PubMed]
58. Pucchio, A.; Eisenhauer, E.A.; Moraes, F.Y. Medical students need artificial intelligence and machine learning training. *Nat. Biotechnol.* **2021**, *39*, 388–389. [CrossRef] [PubMed]

cancers MDPI

Supplementary Material: Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers

Jeppe Thagaard, Elisabeth Specht Stovgaard, Line Grove Vognsen, Søren Hauberg, Anders B. Dahl, Thomas Ebstrup, Johan Doré, Rikke Egede Vincent, Rikke Karlin Jepsen, Anne Roslind, Iben Kümler, Dorte Nielsen and Eva Balslev


Excluded after review of pathology reports (n=166)
Microinvasion: 14
Not primary BC: 3
Recurrence: 54
Other malignant disease: 13
Bilateral/multifocal: 21
Neoadjuvant: 51
HER2 inconclusive: 8
Progesterone positive: 2
TNBC 01.01.2004-31.12.2014 (n=600)
HER2 analysis not previously performed (n=76)
HER2-positive: 19
HE-stained slides not available (n=153)
TNBC with HE-stained stained slides available (n=262)
Clinical data missing (n=5)
Study population (n=257)

Supplementary Figure S1. Flowchart of patients included in the study.

Supplementary Table S1. Clinicopathological characteristics of the patient population.

N = 262 **Method**	**Number of patients (%)**
Age	
≤50	66 (25.2)
>50	196 (74.8)
Tumor size	
≤2	108 (41.2)
>2	153 (58.4)
Unknown	1 (0.4)
Tumor type	
Ductal	47 (76)
Lobular	3 (1.1)
Other	60 (22.9)
Nodal status	
0	47 (17.9)
1	18 (6.9)
2	16 (6.1)
Unknown	5 (1.9)
Type of operation	
Mastectomy	100 (38.2)
Lumpectomy	162 (61.8)

Appendix C

Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer

In this appendix, we include:

Thagaard, J., Hauberg, S., Dahl, A., Ebstrup, T., Doré, J., Roslind, A., Nielsen, D., Balslev, E., Salgado, R., ..., & Stovgaard, E.S. (2021) Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer. To be submitted.

cancers

MDPI

Review (to be submitted)

Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer

Jeppe Thagaard [1, 2*], Søren Hauberg [1] , Anders B. Dahl [1] , Thomas Ebstrup [2] , Johan Doré [2], Anne Roslind [3], Dorte Nielsen[4], Eva Balslev [3], Roberto Salgado [5], …, and Elisabeth Specht Stovgaard [3]

[1] Technical University of Denmark, Kgs. Lyngby, Denmark
[2] Visiopharm A/S, Hoersholm, Denmark
[3] Department of Pathology, Herlev and Gentofte Hospital, Herlev, Denmark
[4] Department of Oncology, Herlev and Gentofte Hospital, Herlev, Denmark
[5] GZA, Antwerp, Belgium

* Correspondence: jth@visiopharm.com;

Simple Summary: Recent breakthroughs in the field of machine learning (ML) has had a major impact on the field of pathology, and hold promise to overcome many pitfalls of new emerging biomarkers that otherwise are difficult to incorporate into clinical practice. Our study aimed at identifying common challenges for newly-developed prognostic tools that use machine learning (ML) to score tumor infiltrating lymphocytes (TILs) in breast cancer. We found that several sources causing inconsistent results are similar to manual assessment, and further categorize challenges unique to ML methods into methodological aspects, data challenges, and validation issues to aid future development efforts. We conclude that even though ML assessment of TILs has prognostic potential comparable to manual scoring, it still has common pitfalls general to the field of computational pathology. However, we are confident that these can be overcome with further development and clinical validation.

Abstract: The clinical significance of the tumor-immune interaction in breast cancer (BC) has been under intense investigation, and tumor infiltrating lymphocytes (TILs) have re-emerged as a robust and reasonably reproducible biomarker for patients with triple-negative (estrogen and progesterone negative, HER2 normal expression) (TNBC). However, it is a challenging biomarker to incorporate into clinical practice. Recent efforts to use machine learning for automated evaluation of TILs to address the complexity of manual scoring guidelines show promising results. We review state-of-the-art approaches and identify pitfalls and development challenges that cause discordant cases to manual TIL assessment. The main source of inconsistent cases is the inclusion of false-positive areas or cells driven by performance on certain tissue patterns, or design choices in the computational implementation. Other pitfalls are similar to manual assessment: technical slide issues, and heterogeneity in the spatial distribution of TILs. However, ML assessment can also produce results beyond human capabilities, and how discrepancies like these are settled requires validation considerations. Therefore, to aid in solving these challenges, we also give an in-depth discussion on ML and image analysis aspects, data challenges, and validation issues that need to be considered before reliable computational reporting of sTILs can be incorporated into the routine clinical management of TNBC patients.

Keywords: deep learning; digital pathology; guidelines; image analysis; pitfalls; prognostic biomarker; triple-negative breast cancer; tumor-infiltrating lymphocytes

Citation: Thagaard, J. et al., Pitfalls in Machine Learning-assessment of stromal tumor infiltrating lymphocytes in breast cancer. *Cancers* **2021**, *13*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

1. Introduction

In recent years, the treatment potential, as well as the prognostic and predictive significance of the tumor-immune interaction in breast cancer (BC), has been under intense investigation [1,2]. In this context, tumor infiltrating lymphocytes (TILs) have been shown to be a robust and reasonably reproducible biomarker [3,4,5]. In BC, especially the basal-like or triple-negative (estrogen and progesterone negative, HER2 normal expression) (TNBC) and HER2 positive subsets exhibit a more pronounced tumor-associated immune infiltrate. In the TNBC group, a 10% increment in TILs results in a 17% increase in overall survival (OS) [6]. TILs have also been shown to be predictive of chemotherapy treatment [7,8]. The evaluation of TILs was recommended in the 2019 St Gallen International Breast Cancer Conference for routine diagnostics of TNBC [9], and in Denmark, the evaluation of TILs is now incorporated in national guidelines as an optional item for TNBC diagnostics.

With this increasing emphasis on TILs in both research and diagnostic settings, it is imperative that evaluation is correct and reproducible. The International Immuno-Oncology Biomarker Working Group on Breast Cancer (TIL-WG) has devised a set of guidelines for manual TILs evaluation on hematoxylin and eosin (H&E) stained slides [10].

Whilst this method of evaluating TILs is reproducible among trained pathologists with intraclass correlation coefficients of 0.77-0.94 for discrete cut-off values [11,12], it is a challenging biomarker to evaluate and does require some degree of training, which can be difficult in a busy clinical setting. Additionally, time consumption when adding additional biomarkers to an already challenging workload can be problematic.

Recent breakthroughs in the field of machine learning (ML) have had a major impact on the field of computational pathology [13]. Automated evaluation of TILs using ML technology - also referred to as computational TIL assessment (CTA) – is no exception. ML is a promising solution to many of the issues of visual TIL assessment (VTA), and can potentially lead to a standardized and reliable evaluation of TILs regardless of the level of pathologist training, with being less time-consuming as an added benefit.

Using ML and digital image analysis to analyze immune infiltration is not a new idea, and has been studied sporadically over the last decade [14,15,16,17]. However, several novel approaches have concurrently shown the prognostic potential of deep neural network-based algorithms for this task [18,19,20].

When developing algorithms to evaluate TILs in breast cancer, there are important aspects, challenges, and pitfalls that need to be considered, and new development and research should be performed before ML tools can be implemented into the routine clinical management of breast cancer.

In this review, we take a closer look at the current state of CTA and specifically focus on how some of the same pitfalls for manual assessment [11] impact these ML-based methods. We do this by categorizing inconsistent cases mentioned in recent CTA studies [18,19,20]. In addition, we extend this analysis to the challenges unique to solving some of the pitfalls with automated approaches to TILs evaluation. We categorize our findings into four main topics; (i) general pathology pitfalls, (ii) ML and image analysis aspects, (iii) data challenges, and (iv) validation issues.

2. Background

Briefly, the TIL-WG guidelines distinguish between intratumoral TILs (iTILs) in direct contact with tumor cells, and stromal TILs (sTILs), which are located in the stromal tissue between islands of tumor cells. It is also imperative that areas of necrosis, ductal and lobular carcinoma in situ (DCIS/LCIS), and normal breast tissue are excluded. TILs are defined as mononuclear immune cells, i.e. lymphocytes and plasma cells. The guidelines recommend focusing on sTILs, as evaluation of these is more reproducible [6]. sTILs are then assessed as a percentage area coverage of total stromal tumor area, and the final score is reported as a continuous variable.

The TIL-WG has also presented a report on how computational assessment of TILs could be designed with the recommendation that; *"computational TIL assessment (CTA) algorithms need to account for the complexity involved in TIL-scoring procedures, and to closely follow guidelines for visual assessment where appropriate"* [21]. Several different approaches to CTA exist from more granular approaches, closely mimicking the guidelines recommended by the TIL-WG, to more coarse strategies with methods also varying in level of automation. Here, we mainly focus on the recent works that concurrently have created different CTA algorithms that adhere to the guideline.

Bai et al. (2021)[19] published a QuPath [22] open-source method, where the pathologist manually draws the tumor region and excludes non-invasive epithelium (DCIS/LCIS and normal lobules). To handle H&E stain variability, stain normalization is applied before the cells are segmented using a traditional image analysis algorithm (watershed cell detection). Finally, a model, trained on extracted cellular features, classifies all cells as either tumor, TILs, fibroblast, or other. From this, the algorithm then outputs 5 different quantitative variables of both area percentages of TILs, TIL densities in the drawn region, and a proportional number of TILs relative to the tumor, stromal, and total cell numbers.

Sun et al. (2021)[20] presented a more comprehensive approach. After the manually drawn tumor and exclusion of non-invasive regions, a tissue-level model detects and excludes necrosis automatically to make sure that necrotic cells are not misinterpreted as lymphocytes. From here, cells are directly detected and classified as malignant epithelial cells, TILs, or other cells by a cell-level model. The classified cells are then used to identify the tumor, stroma, and lymphocyte-dense regions using a rule-based system, which then outputs a regional-based quantitative variable of the area coverage of sTIL.

Thagaard et al (2021)[18] proposed a fully automatic algorithm using commercial software (Visiopharm A/S, Hoersholm, Denmark), where a tissue-level model finds the tissue section on the slide and then detects the tumor, non-invasive, stromal, and necrotic regions with no manual interaction. A cell-level model then detects cells as TILs and outputs a quantitative variable of the density of TILs in the tumor-associated stroma.

Other studies have previously proposed alternative metrics [17] or used other stains than H&E [23,24,25]. We found that these methods lack consistency with the VTA guideline and refer the reader to Amgad et al. (2020) [21] for an overview.

The common findings across these studies are that CTA is observed to have a good to excellent agreement with VTA, and more importantly, independently associated with patient outcome confirming that patients with TNBC and high CTA score have a significantly favorable survival. However, all studies also agree that CTA is not a panacea to the drawbacks of VTA, and there is still much research to be done in terms of handling pitfalls, further development, and clinical validation.

3. Common pitfalls between visual and computational assessment

On behalf of the TIL-WG, Koz et al. (2020) previously identified and reported the most common pitfalls when evaluating TILs manually. Some of these pitfalls are also relevant when developing ML approaches, and in the following, these will be discussed, as well as potential pitfalls unique to the automated approaches to TILs evaluation.

3.1. Including wrong area or cells

Failure in one or more models that detects areas or cells is the most common reason for inconsistent CTA results with manual scores. Including the wrong area in the quantification are generally due to failure cases on the tissue-level models. The tissue-level pitfalls include: (i) including TILs around non-invasive structures (DCIS/LCIS and normal lobules [18,20]. (ii) extensive lymphovascular invasion [20]. (iii) including necrotic regions as stroma due to loose appearance and necrotic areas [20]. (iv) tertiary lymphoid structures (TLS) [20]. (v) variability of tumor subtype morphology [18,20], where mucinous, metaplastic, apocrine, papillary, or lobular carcinomas can be a challenge to generalize to if the development data is skewed towards the more frequent ductal carcinoma pattern. Both Sun et al. (2021) and Bai et al. (2021) point to exclusion of these confounding regions being performed manually, and hence being subject to the same pitfalls as fully manual

VTA. In Thagaard et al. (2021), this step is performed automatically, but reports issues around very difficult DCIS regions, not around benign regions or easily distinguishable DCIS regions. Overall, both manual and automatic approaches have the same pitfalls around difficult DCIS regions, however, to what extent this is critical for a computational tool is still unknown.

The cell-level pitfalls, where wrong cells are included, are less common for CTA. Bai et al. (2021) report catastrophic segmentation failure in 1-2% of cases. The main cause is not being able to discriminate between iTILs and sTILs, hence cases with a high proportion of iTILs were excluded from the study. Also, apoptotic figures, neutrophils, and low-grade tumors can cause false-positive TIL detections [19].

For all of the pitfalls mentioned here, the general challenge is due to the lack of performance on the specific task to be solved. In general, the reason for this is mainly twofold; the approach selected to solve the task (see Section 4), and/or data variability used to train the model during development (see Section 5).

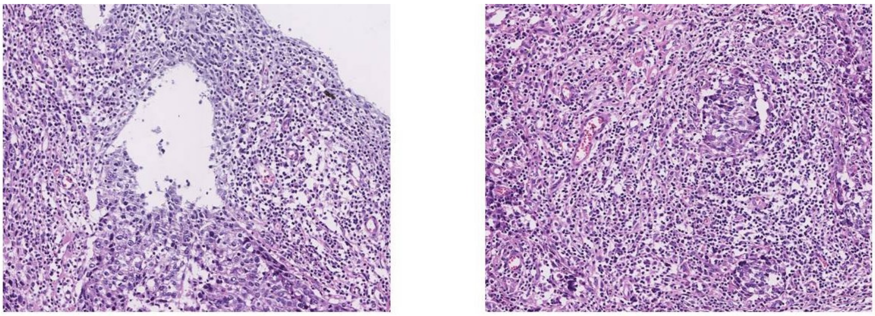


Figure 1. Lymphocyte-dense regions that should be excluded as the inflammation is not necessarily a immune-response to the tumor. Left) is a TLS and right) are lymphocytes surrounding vessels. These areas are reported by Sun et al. (2021) as false-positive areas and result in CTA much higher than VTA. Images from [20] (Supplementary Figure 11 e. and f.).

3.2. *Technical factors*

Furthermore, technical slide-related issues which are also a common challenge for VTA [11], also impact any CTA approach. Artifacts such as out-of-focus areas, pen markings, tissue-folds, and crush artifacts can confuse any tissue- or cell-level model, and cause inaccurate quantification results, e.g. poor sectioning can cause false-negative TILs, hence producing an underestimated TIL density score [18].

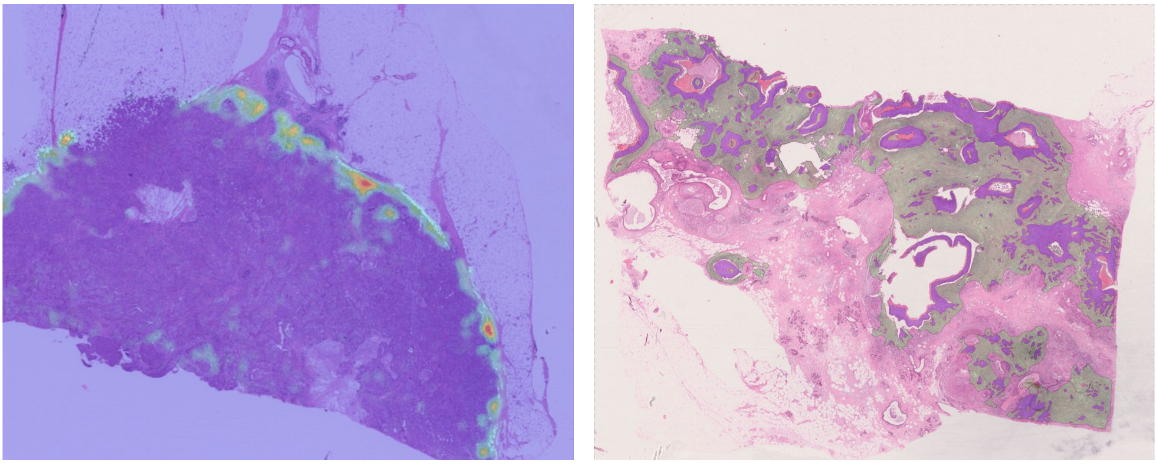
Scanner variability between different manufactures of slide scanners might be an issue when comparing multi-institutional results between cohorts due to the lack of standardization of acquisition parameters. The extend of this issue on CTA is yet to be investigated in depth [19], but for other pathology applications such as prostate cancer detection, results can deteriorate on data from different institutions and scanning systems. We expect the same applies to CTA algorithms. Similarly, both inter- and intra-site variation in H&E staining can contribute to differences between cohorts [19,20] similar to other applications [26].

There are two main approaches to combat these technical factors. First, they can be handled manually or by a separate model, e.g. excluding out-of-focus [27] or folds [28] in a pre-analysis step. Or secondly, incorporating more variability into the dataset used to develop CTA. The latter targets variability in scanning and staining. We cover key aspects of these issues in Section 5.

3.3. *Heterogeneity in sTIL distribution*

One of the pitfalls that cause the highest manual interobserver variation is increased sTILs at the leading edge compared to the central tumor [11], see Figure 2a. It is also mentioned by recent CTA studies [18,20] when comparing CTA to VTA. The increased sTIL density at the leading edge versus central tumor can result in a lower the CTA score as the immune-deserted stromal region in the central tumor will contribute to a larger stroma area quantification than estimated by a manual assessor. On the contrary, if the stroma is scarce in the central tumor, it would be mostly the high density margin that contribute to the overall score.

In general, the definition of the tumor-associated stroma, i.e. the area, which TILs should be scored, is not strictly defined in the manual guideline. Sometimes, there are larger stroma areas within the tumor core, but the exact spatial distance of stromal TIls adjacent to tumor cells nests versus those stromal TILs that are not adjacent to the tumor cell nests is not known, yet. See Figure 2b Similar, the lack of a quantitative definition of a hotspot can be a challenge to adhere to, i.e. which degree should a region be to be considered too TILs-dense for inclusion. This can lead to discrepancies between VTA and CTA [18], but it all depends on the implementation of this into the CTA (Section 4), and the way it is validated (Section 6).



(**a**) Increased sTILs at the leading edge (**b**) Too much stroma included

Figure 2. Examples of discrepant cases from Thagaard et al. (2021). (a) The tumor growas irregularly with small tumor nets between larger invasive tumor areas. In these cases, the CTA included more stroma than VTA, which results in lower sTIL density score (larger denominator) compared to the manual score. (b) A case of high sTIL density at the tumor margin compared to central area. As the stroma is scarce inside the tumor, the sTIL density is reported to very high as mostly the margin contribute to the score.

3.4. Moving beyond human capabilities

Some of the above discrepancies might be eliminated with improved algorithms, such as a finer stroma outlining (see Section 4.2) [20]. However, when the CTA is very precise in its tissue-level outlining as in Thagaard et al. (2021), new pitfalls might arise. Specifically, CTA can include very small areas of stroma within the tumor nests which a pathologist might consider too small, but there are no specific rules on how small a stroma area can be. It can lead to both higher or lower TIL estimation than a manual score if these areas include many TILs (larger TIL count) or do not include TILs (larger stroma area), respectively. When CTA can be more precise than what a pathologist would ever do in practice, it will lead to discrepant cases. However, it again comes down to the validation approach to settle these, which we discuss more in section 6.

4. Image analysis challenges when adhering to a clinical guideline

Many of the pitfalls, mentioned in the previous section, can be contributed to which image analysis approach is used to implement the rules of the VTA guideline. The design choices made here will affect the results. However, in most existing histology-based biomarkers, the current gold reference of scoring is manual expert assessment such as the interpretation guideline for scoring PD-L1 [29], HER2 [30], and the VTA guideline [10]. Hence, the computational pathology community always needs to answer the same question first; what strategy do we want to use to translate the rules of manual guidelines into something a computer can execute? There are many valid answers to this question, and in this section, we review their pros and cons in regards to CTA.

4.1. There exist different solutions to a computer vision problem

There are 3 main categories of computational approaches, one can consider: 1) classification, 2) object detection, and 3) segmentation. There are also other methods in addition to combinations of the three, but for clarity, we focus on the major ones here and refer to Srinidhi et al. (2021) [31] for a more extensive review of methods. First, classification is the simplest approach and is sometimes also referred to as patch-based approaches in pathology. The purpose of this approach is to take an image patch/field-of-view (FOV) and classify it into one discrete label/class from a predefined set of labels, i.e., an image is turned into a single value. Secondly, object detection is the natural expansion from classification as the purpose is not only to map what object is in an image but also where it is located spatially. The location of the objects of interest is usually marked with a bounding box around the four corners of the object or with a single coordinate in the center of the object. Lastly, segmentation is the pixel-wise classification of objects in the image, i.e., every pixel is associated with a label/class which then creates a semantic understanding of what and where objects are in the image. In contrast to object detection, segmentation allows you to outline/delineate the border of the objects very precisely. See Fig. 1 for a graphical overview. The general consideration when selecting a category for a computer vision problem is to determine the level of precision/resolution needed at inference time, i.e., how coarse can the output be to still adhere to the guideline. There are also differences in terms of training data and validation requirements that we will cover in later sections.

As the TIL-WG guideline states [21], a CTA algorithm should be able to do two things; 1) detect and compartmentalize the tissue section, so 2) the quantification of TILs is performed in the right compartment. For these two problems, different considerations apply.

4.2. Considerations for tissue-level models

To detect and compartmentalize the tissue, object detection is often excluded as this approach is not suited to subdivide complex tissue structures into distinct areas, e.g., highly infiltrating tumor nests. Most CTA algorithms use a classification approach [17,32] or full segmentation [16,33]. The main difference involves the coarseness, where classification is less precise than segmentation, but resolution can differ depending on the overlap of the sliding window.

When using a direct classification of image patches as tumor, stroma, or lymphocyte regions, one patch might contain different tissue components, which makes classification difficult. Moreover, this patch-based approach may not provide detailed, quantitative information on the number or density of TILs. Bai et al (2021) used manual outlining by a pathologist and did not as such discriminate tumor and stroma areas, which caused problems with high iTILs cases as mentioned in section 3.1. Sun et al. (2021) also used manual outlining in combination with a patch-based model to detect and exclude necrosis. They then used the cell-level output (see section 4.3), and empirically defined a tumor area as a patch with more than two tumor cells. As also mentioned by the authors, this sliding window approach produced relatively coarse region boundaries compared to a full segmentation model [16; 18].

By providing models that segment the tissue compartments, allows for generating more detailed quantitative information downstream on the cellular level. Even though segmentation seems to be the obvious choice to handle the tissue-level task of finding the stromal area needed for CTA, the approach also has some drawbacks. First, there might be segmentation artifacts from the sliding window analysis, which is still needed due to the gigapixel size of whole slide images (WSIs). This can also lead to issues when the tiling of the WSIs causes that e.g., one glandular structure is divided into two and analyzed independently, and one part is segmented as invasive tumor and the other part as DCIS. In the naïve setup, the machine learning model only takes one part into account at a time in what is called the receptive view, i.e., what is the part of tissue structure physically that the model sees at each prediction. Such inconsistencies along the edges of each FOV need to be handled, and there exist post-processing strategies to do so. E.g., in Thagaard et al (2021), if two segments of DCIS and invasive tumor regions are touching as one object, the size and shape of the DCIS segment are considered in a logical post-processing step to determine if both segments should be segmented as DCIS or invasive tumor. The most important part is that these events are handled consistently, and in relevance to the clinical guideline. For example, one should rather exclude too much epithelium as DCIS, as there is often a high density of stromal TILs around these preinvasive lesions, whereby false-negative DCIS regions would heavily impact the overall TIL score if not excluded as the guideline states.

4.3. Considerations for cell-level models

In the quantification of TILs, the goal is to output a quantitative number of the number of TILs. This step is usually performed at the same or higher magnification to include the cellular level image features in the model. The main goal is to distinguish mononuclear immune cells from all other cell types – both pathological and host cells. Previous studies have solved this task with all 3 main types of approaches. Janowczyk and Madabushi (2016) used a classification model and small sliding window to obtain the most probable localization of each lymphocyte [34]. A potential drawback of this method is that is computationally inefficient as high precision is dependent on having a high degree of overlapping predictions. More recently, several studies [18, 35, 36] have utilized segmentation models to directly predict the center of all the TILs in a FOV, avoiding the inefficiency issue. Others [20,37] used a popular combinatory method of both object detection and segmentation [38] to obtain the localization and outline of TILs at the same time. There are minor differences between the methods, but the main challenges for the cell-level models relate to the requirement of the training data needed to develop them, which we discuss in the section below.

Another consideration is the definition of the final sTIL score as a quantitative output variable. Recent methods have used different definitions as mentioned in section 2. VTA guideline uses an area coverage approach, which is the most accessible for humans to estimate. However, this introduces a slight size bias towards larger TIL nuclei. Does this mean that a CTA should do the same? We argue that as long as the CTA score quantifies the degree of immune infiltration and is human interpretable then it is a valid score, and validation methods will then settle the most appropriate scoring system. The fact that the recent papers found 7 different output variables that all are associated with survival [18,19,20] is evidence for this. Interestingly though, the VTA guideline explicitly states that sTIL should not be scored as a fraction of TILs compared to other cell populations, but Bai et al. (2021) finds two variables of this type consistently provide better results. This indicates that there might be other ways of creating a CTA, but it could also just be a derivative of the model design of the specific paper.

5. Training data challenges to create robust and generalizable algorithms

The models described above are mostly exclusively built using deep learning, a powerful form of machine learning that, given enough training examples, can learn to recognize complex patterns in pathology samples. We will not review all aspects of this field but refer to [31,39]. However, as the most promising CTA algorithms use deep learning, we will in this section cover one of the largest challenges of creating such algorithms; namely the process of obtaining the training data needed to develop them.

5.1. Data variation considerations

The general rule to follow when creating a development dataset is to include the variation which the algorithm is expected to encounter when deployed. Therefore, the requirements strongly depend on the scope of the CTA algorithm, which means what level of generalization is needed. For example, is it a limited single research study, deployment to one single or multiple labs participating in the development (internal generalization), or deployment to a lab outside the development (external generalization)? The answer to this question indicates the possible variation that the CTA algorithm will encounter. The main sources of variation originate from the lack of standardization within pathology.

Variability across pathology laboratories in preanalytical (e.g. fixation, sectioning, etc.) and analytical variables (e.g. H&E staining protocol, scanner type, etc.) causes distributional shifts in the image data. Studies [40] have investigated the impact of such variables, and methods to normalize and/or increase variability from both scanners [40,41,42] and staining [26,43] are being developed.

Another important factor to consider when curating the dataset is the impact of histology subtype (infiltrating ductal, infiltrating lobular, mucinous, etc.) and the histomorphologic variability that this affects in an increasing data distribution. Even the most powerful computational models such as deep learning will not generalize outside the subtype seen during training [44,46], e.g., one should not expect to deploy a model on lobular carcinomas if the development data only included ductal carcinomas. This aspect sets some requirements on how to source and sample the patient population as one should strive for a balanced and realistic dataset between the subtypes.

Luckily, the potential solution to these issues is straightforward in that simply including the variation into the development dataset is by far the most powerful approach to handle it. However, this is a time-consuming process where the tools to do so efficiently are still not adequate. As we can never collect all the variation that exists, there will be some variation the algorithm will encounter. For this, methods to monitor and alarm for novel classes [44], dataset shifts [45,46], and normalization schemes [26,40,43] can potentially mitigate some of the effects of this pitfall.

5.2. Data labeling considerations

Acquiring an adequate number of manual labels is a critical barrier in computational pathology given the time and effort required from pathologists and others with domain expertise. Several approaches have been proposed to address the need for manual labels in large-scale datasets. The requirements for time, expertise, and methods depend on the model type being trained. For the approaches covered in section 2, the magnitude of investment correlates with the precision of the output required. In general, classification labels are the least expensive to obtain, followed by object detection labels and then segmentation labels due to the number of clicks needed to obtain the labels. For the novel approaches being proposed to address this, the main objective is to limit the need to involve pathologists due to the high cost, the time constraints of clinical practice, and the repetitive nature of annotation work.

The most straightforward label strategy is to rely on manual annotations, and scale the number of annotators by involving a large number of experts. This approach has the advantage of guaranteeing high-quality labels, but it is very expensive. One of the pitfalls of manual labels is also the inter-labeler variability and subjectiveness inherent to pathology. A solution to handle this label variability is to get multiple annotators to label the same data, and then make a consensus label. Amgad et al. [37,47] proposed a crowdsourcing framework for both tissue-level segmentation and cell-level classification, object detection, and segmentation. The aim was both to reduce pathologist effort and model the inter-labeler variability of multiple labelers. They show that multiple non-pathologists (up to 6) are required to match the performance of a senior pathologist. However, the benefit is restricted to annotating predominant and visually distinctive patterns. This means pathologist involvement, and possibly full-scale label effort would be needed to supplement uncommon and difficult classes that require greater expertise.

One of the most important parts of developing a labeled dataset for CTA is consistency and not only the scale of the dataset. This consistency is difficult to adhere to when relying on manual labels. Compared to other fields such as radiology, pathology is unique in terms of creating a ground truth definition. For many applications, we rely on manual experts for ground truth, but we also have the option to use the antibody-antigen specificity of immunohistochemical stains (IHC). Recently, Thagaard et al. (2021) proposed to use multiple labeling schemes to obtain the tissue- and cell-level labels, respectively. The general strategy is to use IHC stains during development to help guide semi-automatic labels transfer onto the primary H&E slide, which means that the models then can be trained and deployed on H&E only. The obvious pitfall of this approach is the need to either do new serial sections and thus use more valuable tissue. Also, relevant for TILs, the cellular information might be lost between two consecutive sections due to the cell size being smaller than the section thickness. Alternatively, the H&E section can be restained with IHC if the expertise is available in the laboratory. Thereby ensuring that the IHC stained lymphocytes can be found in the HE-stained slide. However, even though this approach might seem like extra work during development, Thagaard et al. (2021) showed that the quality and consistency of the labels were higher than manual labels, and one pathologist was only needed to review the labels, decreasing the time and effort needed by pathologists.

As WSIs are gigapixel images, it is intractable to label entire WSIs, especially manually. Therefore, one needs to sample training regions to label. Here, it is important to use the same principle to include data (label) variation, e.g. for any TIL model, regions of both low, medium, and high density of TILs should be included, preferably also with varying degrees of proximity to invasive cells.

Even though many different schemes can be used to optimize both the time and need to involve pathologists, such schemes also have their pitfalls as discussed above. We advise always to develop the labeling strategy together with a pathologist and consider it an iterative process to find errors and inconsistencies that help improve the quality and scale of the training labels.

5.3. Data access and sharing considerations

It is obvious that access to data in its raw form is needed to develop CTA algorithms, however, there are significant challenges to collect a dataset of the right cohort. As not many laboratories are fully digital, i.e. scanning all glass slides on modern WSIs scanner into an image management system (IMS) or picture archiving and communication system (PACS), and even less has digitalized historical glass slides, the physical slides of the included patients need to be identified from the archival systems and digitized. Scanning large retrospective datasets can be very time-consuming as most commercial scanners need to be manually operated, but it also opens up for future research if the dataset is stored after the initial study.

A key aspect of developing successful CTA algorithms is collaborations between partners, that is either between academic centers, or between academia and industry partners. Getting the legal terms and conditions into place such that data can be shared between the partners can be a lengthy process. An important aspect to consider is the need to share clinical non-anonymizable metadata as sharing only fully anonymized data significantly eases this process in terms of data privacy protection regulations and the included requirement on the IT infrastructure and security.

There are successful studies sharing high-quality histology datasets publically under Create Commons (CC) licenses [48,49] either fully public [50] or restricted for non-commercial use [51]. The latter can hider some academic-industry collaborations, hence we recommend that datasets should be released under CC4 if possible. The most used platforms for publically sharing datasets are the Grand Challenges website [52], and the Cancer Genome Atlas Program (TCGA)[53]. Historically, there has been a shortage of publicly available datasets for the development of CTA systems with the TCGA cohort being at least part of the foundation for many CTA studies [17,18,20,37]. However, care should be taken to avoid bias and batch effect implications from public datasets not necessarily created for TILs evaluation [54]. There are recent joint efforts from the FDA and TIL-WG to create datasets for algorithm validation [55], however, there is currently a lack of available development datasets. Although collecting a large number of WSIs is time-consuming, it is a manageable task for many pathology laboratories and medical centers, however, collecting training labels remains a barrier to the scaling of CTA algorithms.

6. Validation challenges when comparing CTA with VTA

Both when developing and especially validating any image analysis model, quantitative metrics on the performance of the different parts of the CTA algorithm need to be evaluated. As previously reviewed by the TIL-WG [21], there are different levels of measuring performance. Briefly, analytical validation (AV) refers to low-level metrics such as accuracy and reproducibility, clinical validation (CV) describes the discrimination of patients into clinical subgroups, and clinical utility measures the overall benefit in a clinical setting. In the following, we will discuss some potential pitfalls in model validation.

6.1. Sub-components of modular systems need different evaluation metrics

For the published methods of CTA, it is clear that to adhere to the guideline, the full algorithm is modular, i.e. multiple models are needed to solve different parts of the guideline. Hence, the AV can be applied to the sub-components, and also to the entire system. As the sub-components can be different model approaches, the exact AV metric needs to capture the aspects of each approach, while also be informative about when a failure of a sub-component will cause a failure of the full CTA.

If any sub-component is a segmentation model (e.g. the tumor, necrosis, and non-invasive tissue-level model), standardized metrics as the F1-score can be used to evaluate the AV of the model. The F1-score can be interpreted as the weighted average of the precision and recall/sensitivity. However, one aspect to consider is if it is used locally on a FOV, and there are no true-positive segments of any class, the F1-score will be evaluated as zero, i.e. potential false positives will not be captured as false positives, negatively impacting the overall F1-score. Another challenge for sub-component AV is the impact of the exact score. A benchmark on the exact model selection does not always exist, hence it is difficult to know if an exact score is sufficient, or any better (or worse) model would impact the full CTA AV and/or CV.

For the evaluation of the TIL model, Dudgeon et al. (2020) [55] propose both a metric and a dataset that might qualify as an FDA Medical Device Tool (MDDT) [56]. The proposed metric is a multi-reader multi-case version of the mean squared error (MSE). Other similar metrics like Spearman rank-based correlation are often also used for the algorithm-to-pathologist comparison [16,18]. One of the pitfalls of using such count-based metrics is that they do not capture if the pathologist and algorithm are counting the same TILs but just compare the sum of TILs. However, the metrics are easy to use and interpret, and they capture the most clinically relevant aspect of the algorithm – the extent of TILs in a certain region.

6.2. Considerations regarding clinical validation and utility

For the AV of the full algorithm, the same metrics can be used for the algorithm-to-pathologist comparison. However, as recently commented by Acs et al. (2021) [57], it is still an open question what the best method to evaluate digitally assessed biomarkers such as CTA for both AV and CV. They point out the paradox of selecting the ground truth for digital pathology in TILs as either the concordance between the pathologist and the computational score, patient outcome, or a combination of both. This also raises the question of the clinical cut-off as there are currently no formal recommendations for a clinically relevant cutoff point for stromal TILs. Not having any stratification of patients into clinically meaningful subgroups using manual VTA, makes the CV more challenging for CTA because any cut-off comparison between VTA and CTA might be arbitrary. Current CTA studies [18,19,20] use various cut-off points used for VTA studies [3,58-60] to discriminate between two groups (TILs-high vs. TILs-low) and find different agreement levels between manual and automated methods at different cut-offs. Sun et al. (2021) find moderate-to-substantial agreement depending on the exact cut-offs, but only slight agreement at a 10% cutoff. On the contrary, Thagaard et al. (2021) find substantial agreement at the same 10% cutoff on a different cohort. Interestingly, Sun et al. (2021) findings suggest the possible need for different TILs cut-off values depending on the cohort and different ethnicities, although no significant difference in TIL distribution was found between Asians and Caucasians. This highlights the difficulties of finding a cut-off for biomarkers, in general, which still involve a high degree of uncertainty [Acs et al., 2021]. On the contrary, both studies find that the automated CTA score as a continuous variable is associated with the primary end-points of disease-free survival (DFS) and overall survival (OS). Hence, this suggests that TILs could be better integrated into prognostic modeling containing existing clinical variables such as age, LNS, tumor size, and tumor type. This would remove the need to determine a cut-off, also in different ethnicities. It remains an active field in progress, and there are still aspects that we don’t know, yet.

7. Discussion

Current state-of-art CTA algorithms suggest that stromal TILs can be assessed computationally, and represent a crucial prognostic and predictive factor for TNBC in line with previous VTA findings [6]. It is also clear that there are different approaches to create a CTA algorithm without any conclusion on which constitutes the optimal approach. However, irrespectively of methodology, it is clear that many of the same pitfalls for VTA [11] also cause potential problematic variability for CTA. Whether this has any impact on the clinical validation of CTA is yet unknown as it very much depends on how these algorithms will end up being validated. As stated by Arcs et al (2021), the TIL-WG is currently organizing a grand challenge using phase 3 clinical trial data which is a crucial step to validate any CTA algorithm. The hope is that this will answer many of the questions related to the importance of precision of the algorithmic steps that are currently difficult to evaluate. However, to create highly robust and generalizable CTA algorithms, similar collaborative community-driven initiatives are needed to build labeled development datasets. Currently, there is no public framework or infrastructure to work collaboratively on different labeling strategies ensuring that CTA algorithms can identify and handle all histological components including DCIS within the invasive tumor, fibrosis, hyalinization, and a larger number of granulocytes mentioned as confounding factors in the VTA guideline. We address the lack on this part to rigid institutional requirements governing privacy protection, and that there is currently no easy and practical way of build combined versioned datasets of standardized WSI and label formats.

Another important but unsettled aspect is the human-algorithm interaction within a day-to-day clinical workflow. Should the pathologist be required to open a case and manually draw or edit regions, send the cases for analysis and wait for it to finalize the analysis? Or should the algorithm be automated so the case can be analyzed based on slide metadata directly after scanning, meaning that the case is already analyzed when the pathologist opens the case for the first time? It is yet to been seen which and how different implementations can be optimized to augment and not disrupt the current workflow. Similarly, we still do not know how to best present the quantitative results of CTA, e.g. a precise count of TILs per mm^2 or something closer to an area fraction. Findings also suggest that a dichotomous score of both computationally and manual measurement predicts outcome better than either variable alone [20]. This might affect whether or not the CTA should provide the primary score or work as a secondary reader on difficult cases. There are still many unknowns, but work is in progress on many of the challenges as we look into an exciting future. Until we have more answers to these questions, the diagnostic responsibility still resides with the pathologist and will probably continue to be in any foreseeable future regardless. The main question is then how can we provide tools to pathologists sooner that assist and provide the opportunity to deliver more standardized care for the patients?

We hope that by highlighting the specific pitfalls in using machine learning for sTIL assessment, future developments and collaborations will be armed to find the answers needed to ensure reliable computational reporting of sTILs with the end goal of progressing it into the routine clinical management of breast cancer.

Supplementary Materials: N/A

Author Contributions: Conceptualization, J.T., S.H., A.D., T.E., J.D., R.S., and E.S.; methodology, J.T., E.S.; writing—original draft preparation, J.T., and E.S.; writing—review and editing, All; visualization, J.T.; supervision, S.H., A.D., T.E., J.D., and E.B; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark, grant number 8053-00008B.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable**.**

Data Availability Statement: Not applicable**.**

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: J.T. and T.E. are employees of Visiopharm A/S. J.D. is the co-founder of Visiopharm A/S. All other authors declare no conflict of interest.

References

1. Bates GJ, Fox SB, Han C, Leek RD, Garcia JF, Harris AL, et al. Quantification of regulatory T cells enables the identification of high-risk breast cancer patients and those at risk of late relapse. *J Clin Oncol*. **2006** 24:5373–80. doi: 10.1200/JCO.2006.05.9584
2. Wang M, Zhang C, Song Y, Wang Z, Wang Y, Luo F, et al. Mechanism of immune evasion in breast cancer. Onco Targets *Ther.* **2017** 10:1561–73. doi: 10.2147/OTT.S126424
3. Savas, P. et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nat. Rev. Clin. Oncol*. **2016** 13, 228–241.
4. Hammerl D, Smid M, Timmermans AM, Sleijfer S, Martens JWM, Debets R. Breast cancer genomics and immuno-oncological markers to guide immune therapies. Semin Cancer Biol. 2018;52(Pt 2):178-188. doi:10.1016/j.semcancer.2017.11.003
5. Hudeček J, Voorwerk L, van Seijen M, et al. Application of a risk-management framework for integration of stromal tumor-infiltrating lymphocytes in clinical trials. NPJ Breast Cancer. 2020;6:15. Published 2020 May 12. doi:10.1038/s41523-020-0155-1
6. S. Loi et al., "Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers," J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 37, no. 7, pp. 559–569, 01 2019, doi: 10.1200/JCO.18.01010.
7. Liang, H., Li, H., Xie, Z., Jin, T., Chen, Y., Lv, Z., Tan, X., Li, J., Han, G., He, W., Qiu, N., Jiang, M., Zhou, J., Xia, H., Zhan, Y., Cui, L., Guo, W., Huang, J., Zhang, X., & Wu, Y. L. Quantitative multiplex immunofluorescence analysis identifies infiltrating PD1+ CD8+ and CD8+ T cells as predictive of response to neoadjuvant chemotherapy in breast cancer. *Thoracic cancer*, **2020** 11(10), 2941–2954. https://doi.org/10.1111/1759-7714.13639
8. Russo, L., Maltese, A., Betancourt, L., Romero, G., Cialoni, D., De la Fuente, L., Gutierrez, M., Ruiz, A., Agüero, E., & Hernández, S. Locally advanced breast cancer: Tumor-infiltrating lymphocytes as a predictive factor of response to neoadjuvant chemotherapy. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, **2019** 45(6), 963–968. https://doi.org/10.1016/j.ejso.2019.01.222
9. C. Morigi, "Highlights of the 16th St Gallen International Breast Cancer Conference, Vienna, Austria, 20-23 March 2019: personalised treatments for patients with early breast cancer," Ecancermedicalscience, vol. 13, p. 924, 2019, doi: 10.3332/ecancer.2019.924.
10. Salgado, R.; Denkert, C.; Demaria, S.; Sirtaine, N.; Klauschen, F.; Pruneri, G.; Wienert, S.; Van den Eynden, G.; Baehner, F.L.; Penault-Llorca, F.; et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: Recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **2015**, 26, 259–271.
11. Kos Z, Roblin E, Kim RS, et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. NPJ Breast Cancer. **2020**;6:17. Published 2020 May 12. doi:10.1038/s41523-020-0156-0
12. O'Loughlin M, Andreu X, Bianchi S, et al. Reproducibility and predictive value of scoring stromal tumour infiltrating lymphocytes in triple-negative breast cancer: a multi-institutional study. Breast Cancer Res Treat. 2018;171(1):1-9. doi:10.1007/s10549-018-4825-8
13. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021;27(5):775-784. doi:10.1038/s41591-021-01343-4
14. Basavanhally AN, Ganesan S, Agner S, et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. IEEE Trans Biomed Eng. 2010;57(3):642-653. doi:10.1109/TBME.2009.2035305
15. Yuan Y, Failmezger H, Rueda OM, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling [published correction appears in Sci Transl Med. 2012 Oct 24;4(157):161er6]. Sci Transl Med. 2012;4(157):157ra143. doi:10.1126/scitranslmed.3004330
16. Amgad M, Sarkar A, Srinivas C, et al. Joint Region and Nucleus Segmentation for Characterization of Tumor Infiltrating Lymphocytes in Breast Cancer. Proc SPIE Int Soc Opt Eng. 2019;10956:109560M. doi:10.1117/12.2512892

17. Saltz J, Gupta R, Hou L, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep*. **2018**;23(1):181-193.e7. doi:10.1016/j.celrep.2018.03.086
18. Thagaard J, Stovgaard ES, Vognsen LG, et al. Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers. Cancers (Basel). 2021;13(12):3050. Published 2021 Jun 18. doi:10.3390/cancers13123050
19. Bai Y, Cole K, Martinez-Morilla S, et al. An Open Source, Automated Tumor Infiltrating Lymphocyte Algorithm for Prognosis in Triple-Negative Breast Cancer [published online ahead of print, 2021 Jun 4]. Clin Cancer Res. 2021;clincanres.0325.2021. doi:10.1158/1078-0432.CCR-21-0325
20. Sun P, He J, Chao X, et al. A Computational Tumor-Infiltrating Lymphocyte Assessment Method Comparable with Visual Reporting Guidelines for Triple-Negative Breast Cancer. EBioMedicine. 2021;70:103492. doi:10.1016/j.ebiom.2021.103492
21. Amgad M, Stovgaard ES, Balslev E, et al. Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group. NPJ Breast Cancer. **2020**;6:16. doi:10.1038/s41523-020-0154-2
22. Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, **2017**. https://doi.org/10.1038/s41598-017-17204-5
23. He TF, Yost SE, Frankel PH, et al. Multi-panel immunofluorescence analysis of tumor infiltrating lymphocytes in triple negative breast cancer: Evolution of tumor immune profiles and patient prognosis. PLoS One. **2020**;15(3):e0229955. Published 2020 Mar 9. doi:10.1371/journal.pone.0229955
24. Swiderska-Chadaj Z, Pinckaers H, van Rijthoven M, et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. Med Image Anal. **2019**;58:101547. doi:10.1016/j.media.2019.101547
25. Balkenhol MC, Ciompi F, Świderska-Chadaj Ż, et al. Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics. Breast. 2021;56:78-87. doi:10.1016/j.breast.2021.02.007
26. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019;58:101544. doi:10.1016/j.media.2019.101544
27. Kohlberger T, Liu Y, Moran M, et al. Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection. J Pathol Inform. 2019;10:39. Published 2019 Dec 12. doi:10.4103/jpi.jpi_11_19
28. Smit, G., Ciompi, F., Cigéhn, M., Bodén, A., van der Laak, J., & Mercan, C. Quality control of whole-slide images through multi-class semantic segmentation of artifacts. 2020. In MIDL.
29. Guo, H., Ding, Q., Gong, Y. et al. Comparison of three scoring methods using the FDA-approved 22C3 immunohistochemistry assay to evaluate PD-L1 expression in breast cancer and their association with clinicopathologic factors. Breast Cancer Res 22, 69 (2020). https://doi.org/10.1186/s13058-020-01303-9
30. Wolff AC, Hammond MEH, Allison KH, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. J Clin Oncol. 2018;36(20):2105-2122. doi:10.1200/JCO.2018.77.8738
31. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: A survey. Med Image Anal. 2021;67:101813. doi:10.1016/j.media.2020.101813
32. Le H, Gupta R, Hou L, et al. Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer. Am J Pathol. **2020**;190(7):1491-1504. doi:10.1016/j.ajpath.2020.03.012
33. Abe N, Matsumoto H, Takamatsu R, et al. Quantitative digital image analysis of tumor-infiltrating lymphocytes in HER2-positive breast cancer. Virchows Arch. 2020;476(5):701-709. doi:10.1007/s00428-019-02730-6
34. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016;7:29. Published 2016 Jul 26. doi:10.4103/2153-3539.186902
35. Lu Z, Xu S, Shao W, et al. Deep-Learning-Based Characterization of Tumor-Infiltrating Lymphocytes in Breast Cancers From Histopathology Images and Multiomics Data. JCO Clin Cancer Inform. 2020;4:480-490. doi:10.1200/CCI.19.00126
36. Chen, J. & Srinivas, C.. Automatic Lymphocyte Detection in H&E Images with Deep Neural Networks., 2016
37. Amgad, M., Atteya, L. A., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., ... & Cooper, L. A.. NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. arXiv preprint arXiv:2102.09099. 2021
38. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. PP. 1-1. 10.1109/TPAMI.2018.2844175
39. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88. doi:10.1016/j.media.2017.07.005
40. Swiderska-Chadaj Z, de Bel T, Blanchet L, et al. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. Sci Rep. 2020;10(1):14398. Published 2020 Sep 1. doi:10.1038/s41598-020-71420-0
41. Zarella MD, Bowman D, Aeffner F, et al. A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association. Arch Pathol Lab Med. 2019;143(2):222-234. doi:10.5858/arpa.2018-0343-RA
42. Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019;249(3):286-294. doi:10.1002/path.5331
43. de Bel T, Bokhorst JM, van der Laak J, Litjens G. Residual cyclegan for robust domain transformation of histopathological tissue slides. Med Image Anal. 2021;70:102004. doi:10.1016/j.media.2021.102004

44. Linmans, J., van der Laak, J. &; Litjens, G.. (2020). Efficient Out-of-Distribution Detection in Digital Pathology Using Multi-Head Convolutional Neural Networks. Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR, 2020. 121:465-478
45. Stacke, K., Eilertsen, G., Unger, J., Lundstr ¨om, C.: A closer look at domain shiftfor deep learning in histopathology. arXiv preprintarXiv:1909.11575. 2019
46. Thagaard, J., et al. Can you trust predictive uncertainty under real dataset shifts in digital pathology? MICCAI. 2020:824-833). Springer. https://doi.org/10.1007/978-3-030-59710-8_80
47. Amgad M, Elfandy H, Hussein H, et al. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics. 2019;35(18):3461-3467. doi:10.1093/bioinformatics/btz083
48. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Available online: https://creativecommons.org/publicdomain/zero/1.0/ (accessed on 31 August 2021).
49. Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Available online: https://creativecommons.org/licenses/by-nc-sa/4.0/ (accessed on 31 August 2021).
50. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Gigascience. 2018;7(6):giy065. doi:10.1093/gigascience/giy065
51. Prostate cANcer graDe Assessment (PANDA) Challenge. Available online: https://www.kaggle.com/c/prostate-cancer-grade-assessment (accessed on 31 August 2021).
52. Grand Challenge. Available online: https://grand-challenge.org/ (accessed on 31 August 2021).
53. The Cancer Genome Atlas Program. Available online: https://www.cancer.gov/tcga (accessed on 31 August 2021).
54. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021;12(1):4423. Published 2021 Jul 20. doi:10.1038/s41467-021-24698-1
55. Dudgeon, Sarah N., et al. "A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study." 2020. arXiv preprint arXiv:2010.06995
56. Qualification of Medical Device Development Tools. Available online: https://www.fda.gov/media/87134/download (accessed on 31 August 2021).
57. Acs B, Salgado R, Hartman J. What do we still need to learn on digitally assessed biomarkers?. EBioMedicine. 2021;70:103520. doi:10.1016/j.ebiom.2021.103520
58. Stanton SE, Disis ML. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. J Immunother Cancer 2016 Dec;4(1):1–7
59. Stanton SE, Adams S, Disis ML. Variation in the incidence and magnitude of tumor-infiltrating lymphocytes in breast cancer subtypes: a systematic review. JAMA Oncol 2016 Oct 1;2(10):1354–60.
60. Adams S, Gray RJ, Demaria S, Goldstein L, Perez EA, Shulman LN, Martino S, Wang M, Jones VE, Saphner TJ, Wolff AC. Prognostic value of tumor-infiltrating lympho- cytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. J. Clin. Oncol. 2014 Sep 20;32 (27):2959.
61. Pruneri G, Vingiani A, Bagnardi V, Rotmensz N, De Rose A, Palazzo A, Colleoni AM, Goldhirsch A, Viale G. Clinical validity of tumor-infiltrating lymphocytes analysis in patients with triple-negative breast cancer. Ann. Oncol. 2016 Feb 1;27(2):249– 56

Bibliography

[1] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” *Medical Image Analysis*, vol. 58, p. 101544, 2019.

[2] C. E. DeSantis, F. Bray, J. Ferlay, J. Lortet-Tieulent, B. O. Anderson, and A. Jemal, “International variation in female breast cancer incidence and mortality rates,” *Cancer Epidemiology and Prevention Biomarkers*, vol. 24, no. 10, pp. 1495–1506, 2015.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016,” *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[4] “Breast cancer.” `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`. Accessed: 31-08-2021.

[5] “Invasive ductal carcinoma of the breast.” `https://www.mypathologyreport.ca/breast-invasive-ductal-carcinoma/`. Accessed: 31-08-2021.

[6] C. Perou, T. Sørile, M. Eisen, M. Van De Rijn, S. Jeffrey, C. Ress, J. Pollack, D. Ross, H. Johnsen, L. Akslen, Ø. Fluge, A. Pergammenschlkov, C. Williams, S. Zhu, P. Lønning, A. Børresen-Dale, P. Brown, and D. Botstein, “Molecular portraits of human breast tumours,” *Nature*, vol. 406, pp. 747–752, Aug. 2000.

[7] P. Carter, L. Presta, C. M. Gorman, J. B. Ridgway, D. Henner, W. L. Wong, A. M. Rowland, C. Kotts, M. E. Carver, and H. M. Shepard, “Humanization of an anti-p185her2 antibody for human cancer therapy.,”

Proceedings of the National Academy of Sciences, vol. 89, no. 10, pp. 4285–4289, 1992.

[8] T. O. Nielsen, S. C. Y. Leung, D. L. Rimm, A. Dodson, B. Acs, S. Badve, C. Denkert, M. J. Ellis, S. Fineberg, M. Flowers, H. H. Kreipe, A.-V. Laenkholm, H. Pan, F. M. Penault-Llorca, M.-Y. Polley, R. Salgado, I. E. Smith, T. Sugie, J. M. S. Bartlett, L. M. McShane, M. Dowsett, and D. F. Hayes, "Assessment of Ki67 in Breast Cancer: Updated Recommendations From the International Ki67 in Breast Cancer Working Group," *JNCI: Journal of the National Cancer Institute*, vol. 113, pp. 808–819, 12 2020.

[9] "Fda grants accelerated approval to pembrolizumab for locally recurrent unresectable or metastatic triple negative breast cancer." `https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-accelerated-approval-pembrolizumab-locally/../-recurrent-unresectable-or-metastatic-triple`. Accessed: 31-08-2021.

[10] Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang, W. Luo, B. Liu, Y. Lei, S. Du, A. Vuppalapati, H. Luu, R. Haydon, T.-C. He, and G. Ren, "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis," *Genes Diseases*, vol. 5, 05 2018.

[11] J. Vestjens, M. Pepels, M. de Boer, G. Borm, C. M. van Deurzen, P. van Diest, J. van Dijck, E. Adang, J. Nortier, E. T. Rutgers, C. Seynaeve, M. Menke-Pluymers, P. Bult, and V. Tjan-Heijnen, "Relevant impact of central pathology review on nodal classification in individual breast cancer patients," *Annals of Oncology*, vol. 23, no. 10, pp. 2561–2566, 2012.

[12] D. Denardo, P. Andreu, and L. Coussens, "Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity," *Cancer metastasis reviews*, vol. 29, pp. 309–16, 06 2010.

[13] R. Huss, C. Schmid, M. Manesse, J. Thagaard, and B. Märkl, "Immunological tumor heterogeneity and diagnostic profiling for advanced and immune therapies," *Advances in Cell and Gene Therapy*, vol. 4, 06 2021.

[14] "What are tils and why are they important ?." `https://www.tilsinbreastcancer.org/what-are-tils/`. Accessed: 31-08-2021.

[15] W. E. Sistrunk and W. C. MacCarty, "Life expectancy following radical amputation for carcinoma of the breast," *Annals of Surgery*, vol. 1, no. 75, pp. 61–69, 1922.

[16] S. Aaltomaa, P. Lipponen, M. Eskelinen, V.-M. Kosma, S. Marin, E. Alhava, and K. Syrjänen, "Lymphocyte infiltrates as a prognostic variable in

female breast cancer," *European Journal of Cancer*, vol. 28, no. 4, pp. 859–864, 1992.

[17] R. Simon, S. Paik, and D. Hayes, "Simon rm, paik s, hayes dfuse of archived specimens in evaluation of prognostic and predictive biomarkers. j natl cancer inst 101: 1446-1452," *Journal of the National Cancer Institute*, vol. 101, pp. 1446–52, 10 2009.

[18] C. Denkert, S. Darb-Esfahani, B. Lederer, B. Heppner, K. Weber, J. Budczies, J. Huober, F. Klauschen, J. Furlanetto, W. Schmitt, J.-U. Blohmer, T. Karn, B. Pfitzner, S. Kümmel, K. Engels, A. Schneeweiss, A. Hartmann, A. Noske, and S. Loibl, "Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy," *The Lancet Oncology*, vol. 19, 12 2017.

[19] S. Loi, D. Drubay, S. Adams, G. Pruneri, P. A. Francis, M. Lacroix-Triki, H. Joensuu, M. V. Dieci, S. Badve, S. Demaria, R. Gray, E. Munzone, J. Lemonnier, C. Sotiriou, M. J. Piccart, P.-L. Kellokumpu-Lehtinen, A. Vingiani, K. Gray, F. Andre, C. Denkert, R. Salgado, and S. Michiels, "Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers," *Journal of Clinical Oncology*, vol. 37, no. 7, pp. 559–569, 2019. PMID: 30650045.

[20] C. Morigi, "Highlights of the 16th st gallen international breast cancer conference, vienna, austria, 20–23 march 2019: personalised treatments for patients with early breast cancer," *ecancermedicalscience*, vol. 13, 04 2019.

[21] J. Bancroft and M. Gamble, "Theory and practice of histological techniques.6th edition," *Churchill Livingstone Elsevier*, pp. 126–127, 01 2008.

[22] "Immunohistochemical staining methods." `https://www.agilent.com/cs/library/technicaloverviews/public/08002_ihc_staining_methods.pdf`. Accessed: 31-08-2021.

[23] "Dicom wg-26: Pathology." `https://www.dicomstandard.org/activity/wgs/wg-26`. Accessed: 31-08-2021.

[24] "Assessment of staining quality." `https://patents.google.com/patent/US10209165B2/`. Accessed: 31-08-2021.

[25] G. Smit, F. Ciompi, M. Cigéhn, A. Bodén, J. van der Laak, and C. Mercan, "Quality control of whole-slide images through multi-class semantic segmentation of artifacts," 2021.

[26] H. Marble, R. Huang, S. Dudgeon, A. Lowe, M. Herrmann, S. Blakely, M. Leavitt, M. Isaacs, M. Hanna, A. Sharma, J. Veetil, P. Goldberg, J. Schmid, L. Lasiter, B. Gallas, E. Abels, and J. Lennerz, "A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients," *Journal of Pathology Informatics*, vol. 11, p. 22, 08 2020.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278 – 2324, 12 1998.

[28] D. Kingma and M. Welling, "Auto-encoding variational bayes," 12 2014.

[29] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021.

[30] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *Proceedings of The 33rd International Conference on Machine Learning*, 06 2015.

[31] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?," *Structural Safety*, vol. 31, pp. 105–112, 03 2009.

[32] T. Pearce, A. Brintrup, and J. Zhu, "Understanding softmax confidence and uncertainty," 2021.

[33] D. Mackay, "Bayesian neural networks and density networks," *Nuclear Instruments and Methods in Physics Research Section A Accelerators Spectrometers Detectors and Associated Equipment*, vol. 354, 10 1999.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014.

[35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 12 2016.

[36] L. Hansen and P. Salamon, "Neural network ensembles," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, pp. 993 – 1001, 11 1990.

[37] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," 2021.

[38] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," 2017.

[39] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” 2015.

[40] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” 2018.

[41] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, “How does mixup help with robustness and generalization?,” 2021.

[42] A. Lamb, V. Verma, J. Kannala, and Y. Bengio, “Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy,” pp. 95–103, 11 2019.

[43] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, “On mixup training: Improved calibration and predictive uncertainty for deep neural networks,” 2020.

[44] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.

[45] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, “Robustness via curvature regularization, and vice versa,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning,” *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2015, pp. 2901–2907, 04 2015.

[47] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, “Pitfalls of in-domain uncertainty estimation and ensembling in deep learning,” *arXiv preprint arXiv:2002.06470*, 2020.

[48] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *Proceedings of International Conference on Learning Representations*, 2017.

[49] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” 2018.

[50] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” 2020.

[51] W. Liu, X. Wang, J. Owens, and Y. Li, “Energy-based out-of-distribution detection,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 21464–21475, Curran Associates, Inc., 2020.

[52] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, T. Cemgil, S. M. A. Eslami, and O. Ronneberger, "Contrastive training for improved out-of-distribution detection," 2020.

[53] B. Ehteshami Bejnordi, M. Veta, J. P, al, F. Beca, S. Albarqouni, R. Cetin-Atalay, T. Qaiser, I. Serrano Gracia, M. Shaban, A. Kalinovsky, H. Matsuda, S. Seno, K. Kartasalo, and D. Racoceanu, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, pp. 2199–2210, 12 2017.

[54] P. Bandi *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.

[55] G. Litjens *et al.*, "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset," *GigaScience*, vol. 7, no. 6, 2018.

[56] Y. Liu, T. Kohlberger, M. Norouzi, G. Dahl, J. Smith, A. Mohtashamian, N. Olson, L. Peng, J. Hipp, and M. Stumpe, "Artificial intelligence based breast cancer nodal metastasis detection: Insights into the black box for pathologists," *Archives of Pathology & Laboratory Medicine*, vol. 143, no. 7, pp. 859–868, 2018.

[57] D. F. Steiner, R. MacDonald, Y. Liu, P. Truszkowski, J. D. Hipp, C. Gammage, F. Thng, L. Peng, and M. C. Stumpe, "Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer," *The American Journal of Surgical Pathology*, vol. 42, no. 12, p. 1636, 2018.

[58] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016.

[59] "90159, ce-ivd metastasis detection, ai." `https://visiopharm.com/app-center/app/metastasis-detection-ai/`. Accessed: 31-08-2021.

[60] "Automatic classification on patient-level breast cancer metastases." `https://grand-challenge-public-prod.s3.amazonaws.com/evaluation-supplementary/80/46fc579c-51f0-40c4-bd1a-7c28e8033f33/Camelyon17_.pdf`. Accessed: 31-08-2021.

[61] Z. Swiderska, T. Bel, L. Blanchet, A. Baidoshvili, D. Vossen, J. Laak, and G. Litjens, "Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer," *Scientific Reports*, vol. 10, 09 2020.

[62] "Cancer classification." https://training.seer.cancer.gov/disease/categories/classification.html. Accessed: 31-08-2021.

[63] J. Linmans, J. van der Laak, and G. Litjens, "Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks," in *Medical Imaging with Deep Learning*, 2020.

[64] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[65] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *NeurIPS*, pp. 13991–14002, 2019.

[66] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[67] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[68] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[69] B. Settles, "Active learning literature survey," 07 2010.

[70] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. Baehne, F. Penault-Llorca, E. Perez, A. Thompson, F. Symmans, A. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, and S. Loi, "The evaluation of tumor-infiltrating lymphocytes (tils) in breast cancer: Recommendations by an international tils working group 2014," *Annals of Oncology*, vol. 26, 09 2014.

[71] C. Denkert, S. Loibl, A. Noske, M. Roller, B. M. Müller, M. Komor, J. Budczies, S. Darb-Esfahani, R. Kronenwett, C. Hanusch, C. von Törne, W. Weichert, K. Engels, C. Solbach, I. Schrader, M. Dietel, and G. von Minckwitz, "Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer," *Journal of Clinical Oncology*, vol. 28, no. 1, pp. 105–113, 2010. PMID: 19917869.

[72] S. Hendry, R. Salgado, T. Gevaert, P. Russell, T. John, B. Thapa, M. Christie, K. Van de Vijver, M. Estrada, P. Gonzalez Ericsson, M. Sanders, B. Solomon, C. Solinas, G. Van den Eynden, Y. Allory, M. Preusser, J. Hainfellner, G. Pruneri, A. Vingiani, and S. Fox, "Assessing tumor-infiltrating lymphocytes in solid tumors: A practical review for pathologists and proposal for a standardized method from the international immunooncology biomarkers working group," *Advances In Anatomic Pathology*, vol. 24, p. 1, 08 2017.

[73] R. Salgado, C. Denkert, C. Campbell, P. Savas, P. Nuciforo, C. Aura, E. Azambuja, H. Eidtmann, C. Ellis, J. Baselga, M. Piccart-Gebhart, S. Michiels, I. Bradbury, C. Sotiriou, and S. Loi, "Tumor-infiltrating lymphocytes and associations with pathological complete response and event-free survival in her2-positive early-stage breast cancer treated with lapatinib and trastuzumab a secondary analysis of the neoaltto trial," *JAMA Oncology*, vol. 1, 04 2015.

[74] Z. Kos, E. Roblin, R. Kim, S. Michiels, B. Gallas, W. Chen, K. Van de Vijver, S. Goel, S. Adams, S. Demaria, G. Viale, T. Nielsen, S. Badve, W. Symmans, C. Sotiriou, D. Rimm, S. Hewitt, C. Denkert, S. Loibl, and K. El Bairi, "Pitfalls in assessing stromal tumor infiltrating lymphocytes (stils) in breast cancer," *npj Breast Cancer*, vol. 6, 12 2020.

[75] F. Klauschen, K. R. Müller, A. Binder, M. Bockmayr, M. Hägele, P. Seegerer, S. Wienert, G. Pruneri, S. de Maria, S. Badve, S. Michiels, T. O. Nielsen, S. Adams, P. Savas, F. Symmans, S. Willis, T. Gruosso, M. Park, B. Haibe-Kains, B. Gallas, A. M. Thompson, I. Cree, C. Sotiriou, C. Solinas, M. Preusser, S. M. Hewitt, D. Rimm, G. Viale, S. Loi, S. Loibl, R. Salgado, and C. Denkert, "Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning," oct 2018.

[76] M. Amgad, E. S. Stovgaard, E. Balslev, J. Thagaard, W. Chen, S. Dudgeon, A. Sharma, J. K. Kerner, C. Denkert, Y. Yuan, K. AbdulJabbar, S. Wienert, P. Savas, L. Voorwerk, A. H. Beck, A. Madabhushi, J. Hartman, M. M. Sebastian, H. M. Horlings, J. Hudeček, F. Ciompi, D. A. Moore, R. Singh, E. Roblin, M. L. Balancin, M.-C. Mathieu, J. K. Lennerz, P. Kirtani, I.-C. Chen, J. P. Braybrooke, G. Pruneri, S. Demaria, S. Adams, S. J. Schnitt, S. R. Lakhani, F. Rojo, L. Comerma, S. S. Badve, M. Khojasteh, W. F. Symmans, C. Sotiriou, P. Gonzalez-Ericsson, K. L. Pogue-Geile, R. S. Kim, D. L. Rimm, G. Viale, S. M. Hewitt, J. M. S. Bartlett, F. Penault-Llorca, S. Goel, H.-C. Lien, S. Loibl, Z. Kos, S. Loi, M. G. Hanna, S. Michiels, M. Kok, T. O. Nielsen, A. J. Lazar, Z. Bago-Horvath, L. F. S. Kooreman, J. A. W. M. van der Laak, J. Saltz, B. D. Gallas, U. Kurkure, M. Barnes, R. Salgado, and L. A. D. Cooper, "Report on computational assessment of Tumor Infiltrating Lymphocytes from the

International Immuno-Oncology Biomarker Working Group," *npj Breast Cancer*, vol. 6, no. 1, 2020.

[77] M. Grunkin, J. Raundahl, and N. Foged, "Practical considerations of image analysis and quantification of signal transduction ihc staining," *Methods in molecular biology (Clifton, N.J.)*, vol. 717, pp. 143–54, 01 2011.

[78] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[79] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018.

[80] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[81] Z. Swiderska, H. Pinckaers, M. Rijthoven, M. Balkenhol, M. Melnikova, O. Geessink, Q. Manson, M. Sherman, A. Polónia, J. Parry, M. Abubakar, G. Litjens, J. Laak, and F. Ciompi, "Learning to detect lymphocytes in immunohistochemistry with deep learning," *Medical Image Analysis*, vol. 58, p. 101547, 08 2019.

[82] M. Balkenhol, F. Ciompi, Z. Swiderska, R. Loo, M. Intezar, I. Otte-Höller, D. Geijs, J. Lotz, N. Weiss, T. Bel, G. Litjens, P. Bult, and J. Laak, "Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics," *The Breast*, vol. 56, 02 2021.

[83] M. Sun, G. Zhang, H. Dang, X. Qi, X. Zhou, and Q. Chang, "Accurate gastric cancer segmentation in digital pathology images using deformable convolution and multi-scale embedding networks," *IEEE Access*, vol. PP, pp. 1–1, 05 2019.

[84] Y. Bao, J. Zhang, Q. Zhang, J. Chang, D. Lu, and Y. Fu, "Artificial intelligence-aided recognition of pathological characteristics and subtype classification of superficial perivascular dermatitis," *Frontiers in Medicine*, vol. 8, p. 696305, 07 2021.

[85] R. C. Gonzalez and R. E. Woods, *Digital image processing.* Upper Saddle River, N.J.: Prentice Hall, 2008.

[86] A. Moreira and M. Santos, "Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points.," pp. 61–68, 01 2007.

[87] D. Tellez, M. Balkenhol, I. Otte-Höller, R. Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi, "Whole-slide mitosis detection in he breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks," *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 03 2018.

[88] S. Stenman, D. Bychkov, H. Kucukel, N. Linder, C. Haglund, J. Arola, and J. Lundin, "Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, 05 2020.

[89] B. Ács, R. Salgado, and J. Hartman, "What do we still need to learn on digitally assessed biomarkers?," *EBioMedicine*, vol. 70, 08 2021.

[90] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part i: Basic concepts and first analyses," *British journal of cancer*, vol. 89, pp. 232–8, 08 2003.

[91] E. Kaplan and P. Meier, "Nonparametrics estimates for incomplete observations," *Journal of the American Statistical Association*, vol. 53, pp. 457–480, 01 1958.

[92] D. Cox, "Regression models and life table," *Journal of the Royal Statistical Society. Series B*, vol. 34, 01 1972.

[93] M. Bradburn, T. Clark, S. Love, and D. Altman, "Survival analysis part ii: Multivariate data analysis- an introduction to concepts and methods," *British journal of cancer*, vol. 89, pp. 431–6, 09 2003.

[94] E. Millar, L. Browne, J. Beretov, K. Lee, J. Lynch, A. Swarbrick, and P. Graham, "Tumour stroma ratio assessment using digital image analysis predicts survival in triple negative and luminal breast cancer," *Cancers*, vol. 12, p. 3749, 12 2020.

[95] S. Loi, P. Schmid, J. Cortes, D. W. Cescon, E. P. Winer, D. L. Toppmeyer, H. S. Rugo, M. De Laurentiis, R. Nanda, H. Iwata, A. Awada, A. R. Tan, R. Salgado, V. Karantza, P. Jelinic, A. Wang, L. Huang, R. Cristescu, L. Annamalai, J. Yearley, J. Yearley, and S. Adams, "Abstract pd14-07: Association between biomarkers and response to pembrolizumab in patients with metastatic triple-negative breast cancer (mtnbc): Exploratory analysis from keynote-086," *Cancer Research*, vol. 81, no. 4 Supplement, pp. PD14–07–PD14–07, 2021.

[96] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021.

[97] R. Li, J. Yao, X. Zhu, Y. Li, and J. Huang, "Graph cnn for survival analysis on whole slide pathological images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), (Cham), pp. 174–182, Springer International Publishing, 2018.

[98] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. u. A. A. Minhas, "Capturing cellular topology in multi-gigapixel pathology images," *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 05 2020.

[99] J. Saltz et al., "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images," *Cell Reports*, vol. 23, pp. 181–193.e7, apr 2018.

[100] Y. Bai, K. Cole, S. Martinez-Morilla, F. S. Ahmed, J. Zugazagoitia, J. Staaf, A. Bosch, A. Ehinger, E. Niméus, J. Hartman, B. Acs, and D. L. Rimm, "An Open Source, Automated Tumor Infiltrating Lymphocyte Algorithm for Prognosis in Triple-Negative Breast Cancer," *Clinical Cancer Research*, p. clincanres.0325.2021, 2021.

[101] P. Sun, J. He, X. Chao, K. Chen, Y. Xu, Q. Huang, J. Yun, M. Li, R. Luo, J. Kuang, H. Wang, H. Li, H. Hui, and S. Xu, "A Computational Tumor-Infiltrating Lymphocyte Assessment Method Comparable with Visual Reporting Guidelines for Triple-Negative Breast Cancer," *EBioMedicine*, vol. 70, p. 103492, 2021.

[102] O. Ciga, A. L. Martel, and T. Xu, "Self supervised contrastive learning for digital histopathology," 2020.