# Topics in Deep Learning Applied to Histopathology Images


Inaugural-Dissertation zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät der Universität zu Köln

vorgelegt von Juan Ignacio Pisula
aus Santa Rosa, La Pampa, Argentina.

Köln, 21.02.2025

| | |
|---|---|
| **Berichterstattende:** | Prof. Dr. Katarzyna Bożek |
| | Prof. Dr. Achim Tresch |
| **Prüfungsvorsitzender:** | Prof. Dr. Kathrin Möllenhoff |

Tag der mündlichen Prüfung: 12.05.2025

# Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäSS aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

Juan I Pisula et al. “Predicting the HER2 status in oesophageal cancer from tissue microarrays using convolutional neural networks”. In: *British Journal of Cancer* 128.7 (2023), pp. 1369–1376

Juan I Pisula et al. “Neural networks predict the pathological response to neoadjuvant radiochemotherapy in esophageal cancer from primary biopsies”. In preparation. 2025

Juan I Pisula et al. “Explainable, federated deep learning model predicts disease progression risk of cutaneous squamous cell carcinoma”. Pre-print. 2024. URL: https://doi.org/10.1101/2024.08.22.24312403

Juan I Pisula and Katarzyna Bozek. “Efficient WSI classification with sequence reduction and transformers pretrained on text”. In: *Scientific Reports* 15.1 (2025), p. 5612. ISSN: 2045-2322

Juan I Pisula and Katarzyna Bozek. “Fine-tuning a Multiple Instance Learning Feature Extractor with Masked Context Modelling and Knowledge Distillation”. In: *Proceedings of the European Conference on Computer Vision (ECCV), Bio Image Computing Workshop*. 2024

Köln, 21.02.2025

Unterschrift

# Abstract

The dissertation at hand presents the main concepts and results derived from research done as a doctoral student in Kasia Bozek's group *Data Science of Bioimages*, and is presented in this thesis as a collection of scientific papers. The thesis touches on several themes, including the interpretation of deep models, multiple instance and federated learning algorithms, and language modeling. These topics are not studied standalone but boarded from the application of computer vision models in the automatic analysis of histopathology images. Emphasis is put on predictive tasks associated with the medical treatment of two diseases, namely, esophageal adenocarcinoma (EA) and cutaneous squamous cell carcinoma (CSCC), which were my main doctoral projects.

The first of the presented works acts as an introduction to the discipline. It studies the prediction of a pathological grading on microarrays of esophageal tissue stained to reveal the presence of a known biomarker, the protein HER2, to identify good candidates for targeted EA therapy. The approach adopted in this paper is the training of an attention-based multiple instance learning classifier, and the explanation of its decision outputs with the aid of saliency maps. This method is the cornerstone of the analyses done in this thesis, and is refined in further chapters.

The upcoming chapters deal with the more challenging problem of prediction of therapy effectiveness from pre-therapy biopsy images, in two different study cases: the response to neoadjuvant radiotherapy in EA patients, and disease progression of CSCC patients treated by tumor excision. Despite the radical differences in tumor biology and therapy procedures, these two problems share many similarities. First of all, this type of prognosis is not done by healthcare professionals, providing no human baselines, hypotheses, or plausible interpretation of results. In the second place, these tasks lack known biomarkers to look for, therefore the tissue sections are stained to reveal their general microscopic anatomy, providing fewer visual cues to learn from. Lastly, from the image analysis standpoint, both problems can be addressed with the same techniques. These two chapters extend the methodology presented in the first work by employing a transformer model as classifier, and an explainability algorithm that suits this new architecture. Additionally, a new analysis stage is added to investigate the cellular composition of relevant image regions via cell nuclei semantic segmentation, feature engineering, and statistical analysis.

The last two showcased works branch from studying the aforementioned disease-specific applications, and explore visual aspects of learning from bioimages. The first of these chapters investigates the impact of pre-training a transformer model with natural language data before being applied to pathology slide classification, and how the visual information in such images can be summarized into smaller representations. The last work in this dissertation proposes a multiple instance learning algorithm incorporating the fact that coarse patterns of tissue morphology and organization are composed of smaller histological features.

# Zusammenfassung

Die vorliegende Dissertation stellt die wichtigsten Konzepte und Ergebnisse aus der Forschung vor, die als Doktorand in Kasia Bozeks Gruppe *Data Science of Bioimages* durchgeführt wurde, und wird in dieser Arbeit als eine Sammlung wissenschaftlicher Arbeiten präsentiert. Die Dissertation befasst sich mit mehreren Themen, darunter die Interpretation von tiefen Modellen, Algorithmen für mehrstufiges und föderiertes Lernen sowie Sprachmodellierung. Diese Themen werden nicht unabhängig voneinander untersucht, sondern sind eingebettet in die Anwendung von Computer-Vision-Modellen bei der automatischen Analyse von histopathologischen Bildern. Der Schwerpunkt liegt auf prädiktiven Aufgaben im Zusammenhang mit der medizinischen Behandlung von zwei Krankheiten, nämlich dem Adenokarzinom der Speiseröhre (EA) und dem Plattenepithelkarzinom der Haut (CSCC), die meine wichtigsten Promotionsprojekte waren.

Die erste der vorgestellten Arbeiten dient als Einführung in diese Disziplin. Sie befasst sich mit der Vorhersage einer pathologischen Einstufung auf Mikroarrays von gefärbtem Speiseröhrengewebe, um das Vorhandensein eines bekannten Biomarkers, des Proteins HER2, aufzudecken und gute Kandidaten für eine gezielte EA-Therapie zu identifizieren. Der in dieser Arbeit verfolgte Ansatz ist das Training eines aufmerksamkeitsbasierten Klassifikators mit mehreren Instanzen und die Erklärung seiner Entscheidungsergebnisse mit Hilfe von Salienzkarten. Diese Methode ist der Eckpfeiler der in dieser Arbeit durchgeführten Analysen und wird in weiteren Kapiteln verfeinert.

Die folgenden Kapitel befassen sich mit dem schwierigeren Problem der Vorhersage der Therapiewirksamkeit anhand von Biopsiebildern vor der Therapie in zwei verschiedenen Studienfällen: dem Ansprechen auf eine neoadjuvante Strahlentherapie bei EA-Patienten und dem Fortschreiten der Erkrankung bei CSCC-Patienten, die durch Tumorexzision behandelt werden. Trotz der radikalen Unterschiede in der Tumorbiologie und den Therapieverfahren weisen diese beiden Probleme viele Gemeinsamkeiten auf. Erstens wird diese Art der Prognose nicht von medizinischem Fachpersonal durchgeführt, so dass es keine menschlichen Grundlagen, Hypothesen oder eine plausible Interpretation der Ergebnisse gibt. Zweitens gibt es bei diesen Aufgaben keine bekannten Biomarker, nach denen man suchen könnte, daher werden die Gewebeschnitte gefärbt, um ihre allgemeine mikroskopische Anatomie zu zeigen, was weniger visuelle Anhaltspunkte bietet, aus denen man lernen kann. Vom Standpunkt der Bildanalyse aus gesehen können beide Probleme mit denselben Techniken angegangen werden. In diesen beiden Kapiteln wird die in der ersten Arbeit vorgestellte Methodik erweitert, indem ein Transformatormodell als Klassifikator und ein Erklärungsalgorithmus verwendet werden, der für diese neue Architektur geeignet ist. Zusätzlich wird eine neue Analysestufe hinzugefügt, um die zelluläre Zusammensetzung relevanter Bildregionen mittels semantischer Segmentierung von Zellkernen, Feature Engineering und statistischer Analyse zu untersuchen.

Die letzten beiden vorgestellten Arbeiten gehen von der Untersuchung der oben genannten krankheitsspezifischen Anwendungen aus und untersuchen visuelle Aspekte des Ler-

nens aus Biobildern. Das erste dieser Kapitel untersucht die Auswirkungen des Vortrainings eines Transformationsmodells mit Daten aus der natürlichen Sprache, bevor es auf die Klassifizierung von Pathologie-Objektträgern angewendet wird, und wie die visuellen Informationen in solchen Bildern in kleineren Darstellungen zusammengefasst werden können. In der letzten Arbeit dieser Dissertation wird ein Lernalgorithmus mit mehreren Instanzen vorgeschlagen, der die Tatsache berücksichtigt, dass grobe Muster der Gewebemorphologie und -organisation aus kleineren histologischen Merkmalen zusammengesetzt sind.

# Abbreviations

**AGEJ**: adenocarcinoma of the gastroesophageal junction

**AI**: artificial intelligence

**CNN**: convolutional neural network

**CSCC**: cutaneous squamous cell carcinoma

**CV**: computer vision

**DL**: deep learning

**EA**: esophageal adenocarcinoma

**H&E**: hematoxylin and eosin

**HER2**: human epidermal growth factor receptor 2

**IHC**: immunohistochemistry

**MIL**: multiple instance learning

**ML**: machine learning

**TMA**: tissue microarray

**WSI**: whole slide image

**XAI**: explainable artificial intelligence

# Contents

Chapter 1

# Introduction

## Histopathology background

Histopathology is the area of medicine that studies the manifestation of diseases via the microscopic examination of tissue samples. It is a fundamental discipline in medical diagnostics, particularly in oncology, where it serves as the gold standard for confirmation of presence or absence of disease, tumor staging and grading, or measurement of disease progression. Histopathology images are the product of digitizing these tissue samples.

Tissue samples taken from the body must first be processed into thin sections that can be mounted on glass slides and stained for microscopic analysis. An important step in the process is tissue staining, as sections are nearly transparent under a microscope, making it impossible to discern pathological features. The main steps in the preparation process are described as follows:

**Fixation.** The first step in tissue preparation is fixation, where tissue specimens obtained from biopsies, surgical resections, or autopsies are immersed in a fixative, typically 10% neutral buffered formalin. This process preserves cellular structures and prevents autolysis and degradation. Fixation times vary depending on tissue type and size but typically range between 6 and 48 hours.

**Grossing and processing.** Before fixation, larger tissue samples undergo grossing, where they are trimmed to appropriate sizes to fit into processing cassettes. Once fixed, tissue samples are processed through a series of steps, including dehydration via graded alcohol solutions, clearing with xylene, and paraffin infiltration. The processed tissues are then embedded in paraffin to create paraffin-embedded tissue blocks, which provide structural support for thin sectioning.

Most commonly, complete tissue sections are analyzed under the microscope or digitized with scanners, however, when large-scale studies require multiple samples to be analyzed under identical conditions, tissue microarrays (TMAs) are employed. TMAs consist of small cylindrical tissue cores arranged in an organized array within a recipient paraffin block, allowing efficient, high-throughput histopathological examination. For their construction, small cylindrical cores (usually 0.62 mm in diameter) are extracted from donor paraffin blocks and arranged into a pre-designed array pattern within a recipient paraffin block. This allows multiple tissue samples to be analyzed on a single slide under uniform conditions.

**Sectioning.** Both standard tissue blocks and TMA blocks are sectioned using a microtome, cutting tissue slices into ultra-thin sections of approximately 45 micrometers. These sections are floated in a warm water bath to remove wrinkles and carefully

mounted onto glass slides. Proper adherence to the slide surface is essential for downstream staining and analysis.

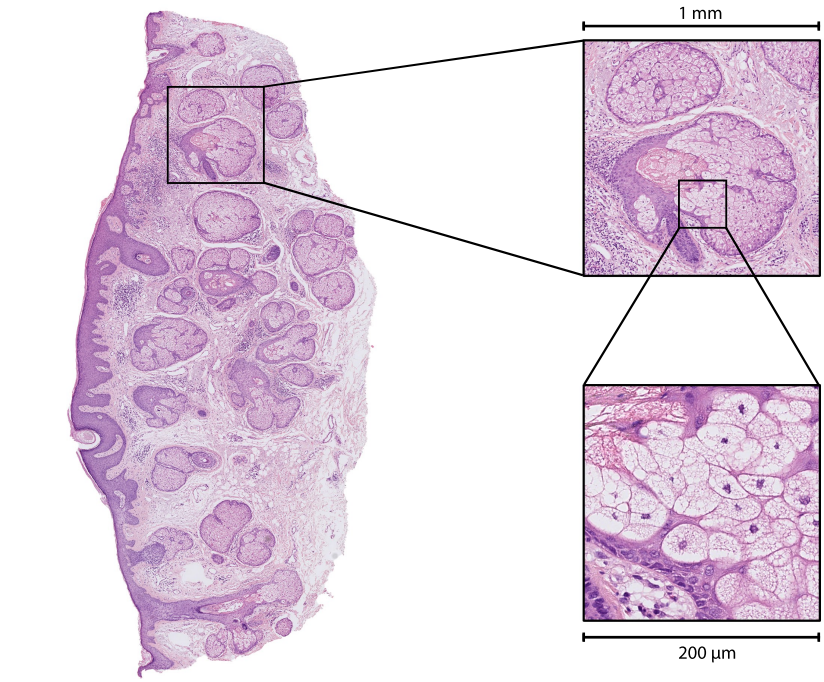
**Deparaffinization and rehydration.** Before staining, the paraffin must be removed from the tissue sections. This is done through deparaffinization, where slides are immersed in xylene to dissolve the paraffin. Following this, the slides are rehydrated by passing them through a graded series of alcohol solutions and rinsing in distilled water to prepare the tissue for staining.

**Staining.** Staining enhances tissue contrast and enables visualization of cellular structures. Two primary staining techniques are employed:

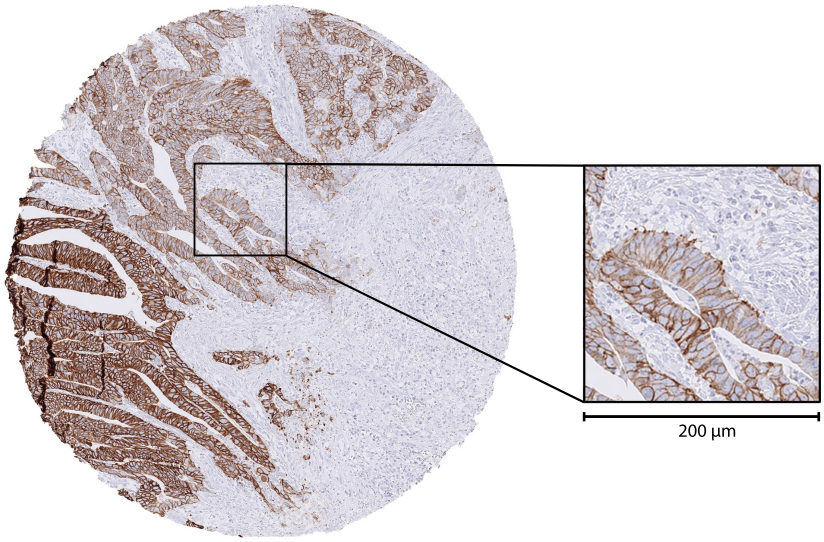
- **Hematoxylin and Eosin (H&E)** staining: This widely used method provides general morphological details. Hematoxylin stains cell nuclei a deep blue or purple, while eosin stains cytoplasmic and extracellular matrix components in varying shades of pink.
- **Immunohistochemistry (IHC)** staining: IHC is used for detecting specific proteins in tissue sections. This involves applying primary antibodies that bind to the target antigen, followed by secondary antibodies conjugated to an enzyme or fluorophore. Enzymatic reactions produce colorimetric signals that highlight specific biomolecules, crucial for diagnosing diseases like cancer by detecting markers such as HER2, Ki-67, and p53. IHC-stained samples require an additional processing step to restore the accessibility of antigens that may have been masked during fixation. This involves heating the slides in a buffer solution or using enzymatic digestion to expose target epitopes for antibody binding.

**Slide drying, coverslipping, and imaging preparation.** Once stained, slides are dried and may be subjected to coverslipping using a mounting medium for long-term preservation. Coverslipping protects the tissue section and enhances optical clarity for microscopic examination. Properly prepared slides are then ready for microscopic examination or digital scanning, enabling computational analysis and remote diagnostics.

Once slides and microarrays are prepared, they must be digitized to enable computational analysis and digital pathology workflows. This process converts the physical glass slides into high-resolution digital images. In the literature, digitized slides are commonly referred as whole slide images (WSIs). The digitization of tissue samples is primarily performed using scanners which operate based on brightfield microscopy principles. In brightfield microscopy, light is transmitted through the stained tissue section, and differences in light absorption create contrast, revealing cellular structures. High-quality objective lenses are used during scanning to achieve the necessary magnification and resolution. Standard magnifications include 20x and 40x, corresponding to approximate pixel resolutions of 0.5 $\mu m$/pixel and 0.25 $\mu m$/pixel, respectively. These resolutions allow pathologists and computational models to analyze minute cellular details effectively.




A




B

**Fig. 1.1:** Examples of histopathology images and stains. **A** A H&E-stained WSI of cutaneous tissue. **B** An IHC-stained TMA core of esophageal tissue, where brown pixels show expression of the HER2 protein.

Slide scanners utilize high-resolution CCD (Charge-Coupled Device) or CMOS (Complementary Metal-Oxide-Semiconductor) sensors to capture the images, technologies used in consumer photographic cameras. Acquisition is done by capturing small, overlapping tissue regions, which are stitched together to form a single image. This method of scanning requires dynamically adjusting the focal plane of the camera to compensate for variations in tissue thickness or glass slide imperfections. Specific details in the acquisition process are vendor-specific, and escape the scope of this dissertation.

Once scanned, digitized slides are stored in multi-resolution pyramidal formats (.tiff, .ndpi, .svs, among others), where lower-resolution versions of the image are generated and stored alongside the full-resolution scan. This enables efficient navigation and zooming without requiring constant access to the highest-resolution data.

## Automatic analysis

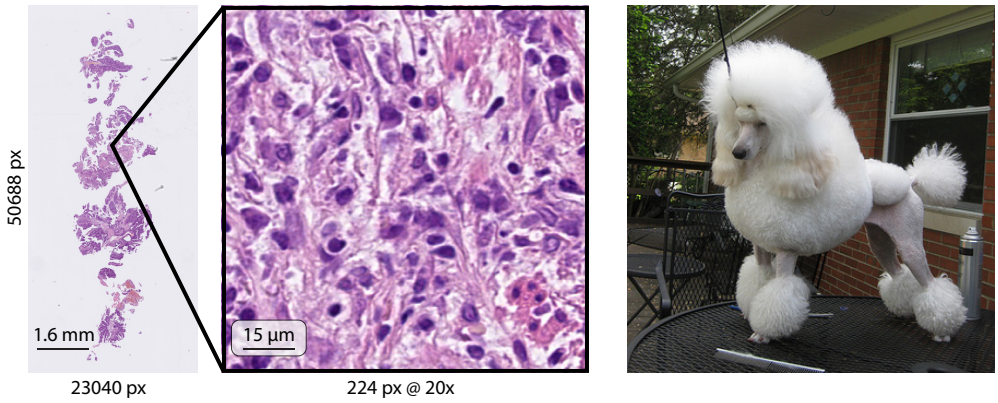



**Fig. 1.2:** Visual differences between histopathology images and natural images. **A** An H&E-stained esophageal slide, comprising several tissue blobs in a clear, distinguishable white background. The zoomed-in region reveals a speckled texture composed of diverse cell nuclei. **B** An image from the well-known ImageNet dataset for visual recognition where the most salient characteristic is the presence of a foreground "object" that belongs to one of the dataset's categories. ImageNet classification models most commonly process images resized to 224 pixels of side.

The availability of digital scans of slides and microarrays allows for the automatic analysis of tissue samples with computer vision (CV) algorithms, including machine and deep learning approaches. Deep models have been proven successful in dense visual tasks such as the detection and segmentation of individual cell nuclei, or global image- and patient-level tasks such as cancer subtype classification or therapy response prediction. This thesis deals with problems of the latter category.

Histopathology imagery is very simple from a technical standpoint. Brightfield microscopy is the most basic form of optical microscopy, and scanners acquire images with

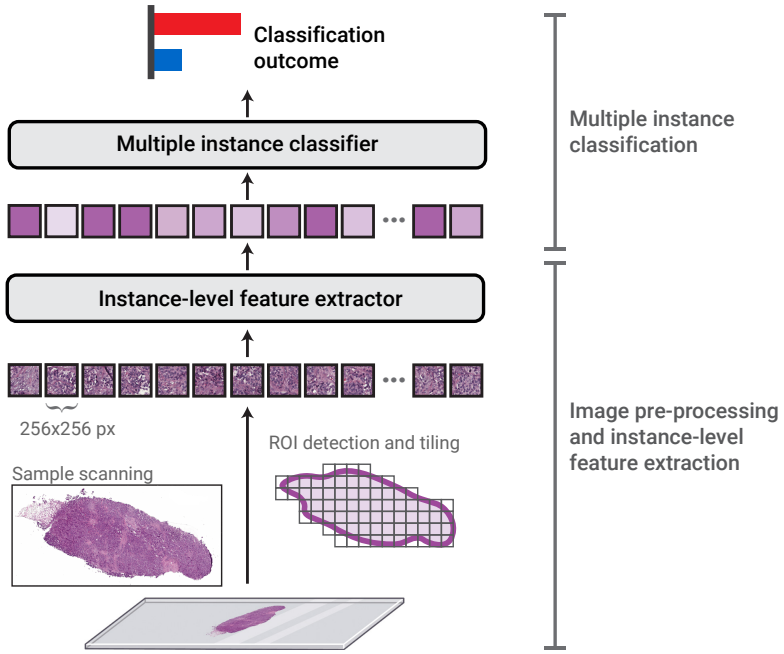



**Fig. 1.3:** A generic 2-stage MIL pipeline. After scanning, the tissue regions of the sample are tessellated into small image patches at a target magnification, and a feature extractor model computes their vector representations. Subsequently, proper MIL classification takes place.

sensors similar to those found in consumer electronics and photographic cameras. However, histopathology images possess distinct traits that differentiate them from other types of imagery.

The most prominent feature is their size: a routine slide scanned at $\times 40$ magnification can easily comprise several gigapixels of resolution. This makes the automatic processing of histopathology images a computational challenge: a single, uncompressed, square slide of 50.000 pixels of side represented in single precision would require 30GB of memory just for its storage. This memory requirement does not take into account the storage required by the image processing algorithm, such as the model's parameters and temporary gradient storage in neural algorithms. Additionally, considering that typical image sizes used in CV algorithms are a couple of hundred pixels per side, just a single slide represents a considerable amount of data to process.

Natural images typically consist of distinct foreground objects (people, animals, things) placed in a background scene. Objects possess high-level visual features (structure, shape, semantic saliency) that the human perception picks up naturally, but are challenging to automatically detect due to low contrast edges, shading and texture gradients, occlusions, that do not create a clear boundary between foreground pixels and back-

ground pixels. On the other hand, histopathology images are composed of tissue blobs, placed on a very distinguishable white background (with no visual information at all), whose main feature is their texture composed by the spatial repetition of cells.

The automatic classification of slide images has adopted the multiple instance learning (MIL) framework. MIL is a type of weakly supervised learning, where labels are assigned to bags of instances, instead of the individual instances themselves. The standard MIL pipeline is illustrated in Fig. 1.3, and is commonly carried out in a pre-processing, offline stage, followed by the training of the MIL classifier itself.

The process begins with the tesselation of the sample into smaller image patches (the instances), at a desired magnification. The choice of magnification is a matter of design, and domain knowledge can aid with this: dense tasks such as identifying slides with metastasis, which manifests itself as a localized cluster of cancer cells, may require 40x or 20x magnification; for tasks such as cancer subtype classification, which can be discriminated by overall tissue features, magnifications of 10x or 5x may suffice.

Empty patches are discarded, and the remaining patches are processed into their (individual) feature vector representations by some already pre-trained deep image model, called the feature extractor. These models follow the standard computer vision architectures, including convolutional neural networks and vision transformers. Feature extractors can be applied following a transfer learning approach, for example, by pre-training them with supervision on a large collection of natural images, as it has been proven that representations learned this way can transfer successfully to other domains. Alternatively, it is also possible to use feature extractors pre-trained with self-supervision on large collections of unlabeled histopathology image data.

The final classification output is produced by the MIL classifier. The main task of this model is to aggregate the complete set of feature vectors of a sample into a single global representation vector, which can then be classified commonly with a learnable linear classification layer. Although the aggregation could be done with any simple pooling operation, like taking the average feature vector of the set, neural attention-based models have been proven successful for this task. Currently, state-of-art MIL classification resides in the use of transformer models, a neural architecture originally developed for natural language processing tasks.

The following chapters explain MIL classification of WSIs and TMAs more comprehensively from various perspectives. The first three chapters focus on the application of MIL classification in predictive tasks associated with the treatment of esophageal adenocarcinoma and cutaneous squamous cell carcinoma. Additionally, these chapters cover topics that exceed the classification itself, namely, the explanation and interpretation of classification results, and collaborative model training. The two chapters that follow study algorithmic aspects of MIL classifiers, as well as the feature extraction of image regions. The dissertation finishes with an overarching discussion.

## References

The Internet Pathology Laboratory for Medical Education. *Histotechniques*. Accessed: 2025-01-01. URL: https://webpath.med.utah.edu/HISTHTML/HISTOTCH/HISTOTCH.html

Ankush Patel et al. "Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review". In: *Journal of Pathology Informatics* 12.1 (2021), p. 50

Metin N Gurcan et al. "Histopathological image analysis: A review". In: *IEEE reviews in biomedical engineering* 2 (2009), pp. 147–171

John Arevalo, Angel Cruz-Roa, and Fabio A González O. "Histopathology image representation for automatic analysis: A state-of-the-art review". In: *Revista Med* 22.2 (2014), pp. 79–91

Chapter 2

# Predicting the HER2 status in oesophageal cancer from tissue microarrays using convolutional neural networks

The following work addresses the classification of IHC-stained TMAs with the ultimate goal of identifying EA patients who qualify for personalized therapy. Tissue cores were stained to reveal the expression of the epidermal growth factor receptor 2 (HER2) protein. HER2 is a well-established cancer biomarker, and its activation is associated with angiogenesis and tumorigenesis. Various solid tumors display HER2 overexpression, and targeted therapy has been shown to improve treatment outcomes. In this study, a neural model is trained to predict the presence of HER2 in two classification tasks: binary classification of HER2 status (positive or negative overexpression) and multi-class classification based on an IHC scoring system that rates HER2 presence. These tasks, the determination of HER2 status and IHC score, are commonly carried out by pathologists (unlike the tasks in the two upcoming chapters) by visually inspecting the IHC-stained tissue cores.

This paper is both the first of the works showcased in this dissertation and the first paper written during my doctoral studies. It serves as an introduction to MIL algorithms applied to histopathology. In this work, the cornerstone method for our approach to analyzing histopathology images is developed: a MIL classifier is trained, and its classification decisions are explained through saliency maps, providing insights into disease mechanisms. The classifier model in this work is a simple attention-based neural network, and its decisions can be explained straightforwardly by visualizing the attention scores each instance receives, presented as a heatmap. The two upcoming chapters build on this approach, employing a different classifier model and a more sophisticated explainability algorithm.

British Journal of Cancer www.nature.com/bjc

ARTICLE OPEN

Check for updates

Molecular Diagnostics

# Predicting the HER2 status in oesophageal cancer from tissue microarrays using convolutional neural networks

Juan I. Pisula[1], Rabi R. Datta[2], Leandra Börner Valdez[2], Jan-Robert Avemarg[3], Jin-On Jung[2], Patrick Plum[2], Heike Löser[4], Philipp Lohneis[4], Monique Meuschke[3], Daniel Pinto dos Santos[5], Florian Gebauer[2], Alexander Quaas[4], Axel Walch[6], Christiane J. Bruns[2], Kai Lawonn[3], Felix C. Popp[2,8] and Katarzyna Bozek[1,7,8]✉



**BACKGROUND:** Fast and accurate diagnostics are key for personalised medicine. Particularly in cancer, precise diagnosis is a prerequisite for targeted therapies, which can prolong lives. In this work, we focus on the automatic identification of gastroesophageal adenocarcinoma (GEA) patients that qualify for a personalised therapy targeting epidermal growth factor receptor 2 (HER2). We present a deep-learning method for scoring microscopy images of GEA for the presence of HER2 overexpression.

**METHODS:** Our method is based on convolutional neural networks (CNNs) trained on a rich dataset of 1602 patient samples and tested on an independent set of 307 patient samples. We additionally verified the CNN's generalisation capabilities with an independent dataset with 653 samples from a separate clinical centre. We incorporated an attention mechanism in the network architecture to identify the tissue regions, which are important for the prediction outcome. Our solution allows for direct automated detection of HER2 in immunohistochemistry-stained tissue slides without the need for manual assessment and additional costly in situ hybridisation (ISH) tests.

**RESULTS:** We show accuracy of 0.94, precision of 0.97, and recall of 0.95. Importantly, our approach offers accurate predictions in cases that pathologists cannot resolve and that require additional ISH testing. We confirmed our findings in an independent dataset collected in a different clinical centre. The attention-based CNN exploits morphological information in microscopy images and is superior to a predictive model based on the staining intensity only.

**CONCLUSIONS:** We demonstrate that our approach not only automates an important diagnostic process for GEA patients but also paves the way for the discovery of new morphological features that were previously unknown for GEA pathology.

*British Journal of Cancer* (2023) 128:1369–1376; https://doi.org/10.1038/s41416-023-02143-y

## BACKGROUND

Gastroesophageal adenocarcinoma (GEA) is the seventh most common cancer worldwide, with an increasing number of cases in the western hemisphere. Despite multimodal therapies with neoadjuvant chemotherapy/chemoradiation before surgery, median overall survival does not exceed 4 years [1–5]. Epidermal growth factor receptor 2 (HER2) encodes a transmembrane tyrosine kinase receptor and is present in different tissues, e.g., epithelial cells, mammary gland, and the nervous system. It is also an important cancer biomarker. HER2 activation is associated with angiogenesis and tumorigenesis. Various solid tumours display HER2 overexpression, and targeted HER2 therapy improves their treatment outcomes [6]. Clinical guidelines for GEA recommend adding Trastuzumab—a monoclonal antibody binding to HER2—to the first-line palliative chemotherapy for HER2-positive cases. HER2 targeting drugs are also currently investigated in the curative therapy for GEA [7].

Accurate testing for the HER2 status is a mandatory prerequisite for the application of targeted therapies. The gold standard for determining the HER2 status is an analysis of the immunohistochemical (IHC) HER2 staining by an experienced pathologist, if necessary followed by an additional in situ hybridisation (ISH). The pathologist examines the immunohistochemistry staining of cancer tissue slides for HER2 and determines the IHC score ranging from 0 to 3. According to expert guidelines [8], the factors determining the score include the staining intensity, the number of connected positive cells, and the cellular location of the staining (Supplemental Table 1). The IHC scores 0 and 1 define patients

[1]Data science of Bioimages Lab, Faculty of Medicine and University Hospital Cologne, Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50931 Cologne, Germany. [2]Department of General, Visceral, Cancer and Transplantation Surgery, University of Cologne, 50937 Cologne, Germany. [3]Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, 07743 Jena, Germany. [4]Institute of Pathology, University of Cologne, 50937 Cologne, Germany. [5]Department of Radiology, University of Cologne, 50937 Cologne, Germany. [6]Research Unit Analytical Pathology, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [7]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany. [8]These authors contributed equally: Felix C. Popp, Katarzyna Bozek. ✉email: k.bozek@uni-koeln.de

Received: 13 May 2022 Revised: 21 December 2022 Accepted: 5 January 2023
Published online: 30 January 2023

with a negative HER2 status that are not eligible for targeted anti-HER2 therapy. An IHC score of 3 designates a positive HER2 status, and these patients receive Trastuzumab. A score of 2 is equivocal. In this case, an additional in situ hybridisation (ISH) assay resolves the IHC score 2 as a positive or negative HER2 status. However, both manual scoring and additional ISH testing are time-consuming and costly.

Automated IHC quantification can support pathologists and is one of the challenges in digital pathology and Convolutional Neural Network (CNN)-based approaches currently offer the highest accuracy in this task [9]. Tewary and Mukhopadhyay using patch-based labelling created a three-level HER2 classifier with an accuracy of 0.93 [10]. Han et al. combined a patch-level classifier with a second one predicting HER2 score of a whole slide image [11] achieving an accuracy of 0.94. The limitation of these methods is the need for patch-level labelling, which is not typically done in clinical evaluation. Annotations of individual patches are not available in clinical datasets and thus require additional manual work while patch-level predictions require developing aggregation strategies to generate a prediction for the entire slide. Additionally, all of the automated methods to date focus on breast tumours, which have high prevalence and offer several large public datasets. HER2 is however an important biomarker in other cancers, notably oesophageal carcinoma.

Here, we ask whether CNNs can directly predict the HER2 status from IHC-stained tissue sections without additional ISH testing. We investigate which image features the neural network learns to make the prediction—whether it uses only the colour intensity or additional morphological features. We explore a large tissue microarray (TMA) with 1602 digitised images stained for HER2. We use this image dataset as a training set to train two different CNN classification models. We test these models on an independent test dataset of 307 TMA images from an unrelated patient group from the same centre. We also further validate the HER2 status prediction accuracy of our approach on a patient cohort from a different clinical centre. If successful, CNNs could assist pathologists in evaluating IHC stainings and, therefore, save time and expenses related to the ISH analysis.

## METHODS

### Tumour sample and image preparation

For training the CNNs, we used a multi-spot tissue microarray (TMA) with 165 tumour cases and a single-spot TMA with 428 tumour cases, as described elsewhere [12]. We additionally prepared an independent single-spot TMA with 307 tumour cases as the test dataset. The test set consisted of tumour cases that occurred at a later time point compared to the training set cases. This dataset construction strategy mimics how such a model would be developed and deployed in a clinical routine. Coincidentally, our test set does not include tumour cases with an IHC score of 1. The multi-spot TMA was composed of eight tissue cores (1.2 mm diameter) of each tumour—four cores punched on the tumour margin and four in the tumour centre. To construct the single-spot TMA, we punched one tissue core per patient from the tumour centre. The cores were transferred to TMA receiver blocks. Each TMA block contained 72 tissue cores. Subsequently, we prepared 4 µm-thick sections from the TMA blocks and transferred them to an adhesive-coated slide system (Instrumedics Inc., Hackensack, NJ).

We used a HER2 antibody (Ventana clone 4B5, Roche Diagnostics, Rotkreuz, Switzerland) on the automated Ventana/Roche slide stainer to perform immunohistochemistry (IHC) on the TMA slides. HER2 expression in carcinoma cells was assessed according to staining criteria listed in Supplemental Table 1. Scores 0 and 1 indicated negative HER2 status, and score 3 indicated positive HER2 status. Immunohistochemical expression evaluation was assessed manually by two pathologists (A.Q. and H.L.) according to [13]. Discrepant results, which occurred only in a small number of samples, were resolved by consensus review. Spots with a score of 2 were analysed by fluorescence ISH to resolve the HER2 status. The ISH analysis evaluated the HER2 gene amplification status using the Zytolight SPEC ERBB2/CEN 17 Dual Probe Kit (Zytomed Systems GmbH, Germany) according to the manufacturer's protocol. A fluorescence microscope (DM5500, Leica, Wetzlar, Germany) with a 63× objective was used for scanning the tumour tissue for amplification hotspots. We counted the signals in randomly chosen areas of homogeneously distributed signals. Twenty tumour cells were evaluated by counting green HER2 and orange centromere-17 (CEN17) signals. The reading strategy followed the recommendations of HER2/CEN17 ratio ≥ 2.0 or HER2 signals ≥ 6.0 for HER2 positive and a HER2/CEN17 ratio <2.0 for HER2-negative samples.

We digitised the slides with a slide scanner (NanoZoomer S360, Hamamatsu Photonics, Japan) with 40-times magnification and used QuPath's [14] TMA dearrayer to slice the digitised slides into individual images (.jpg files, 5468 × 5468 pixels). After discarding corrupted images, this procedure yielded 1281 images for training, 321 validation, and 307 images for testing. The test set is from the same hospital as the train set but was sampled in a time interval disjoint from and following the time interval when the training dataset was collected. This study design not only reflects potential real life clinical scenarios in which incoming patient data is analysed with a model trained on data collected at an earlier time point, but also it follows the guidelines formulated by Kleppe et al. [15].

To study the capability of the CNNs to generalise, we performed a stringent evaluation of the model performance on an external cohort with 653 samples from a different, geographically separate clinical centre [16]. The same antibody was used to perform the HER2 staining, but the slides showed certain deterioration due to aging. Each image was labelled with the IHC score (0, 1, 2, or 3) and the HER2 status (0 or 1) that was determined by the pathologists or by ISH analysis in equivocal cases. This methodology corresponds to the gold standard, and we used this labelling as ground truth.

### Classification models

We implemented a method that allows training neural networks on large images at their original resolution by exploiting weakly supervised Multiple-instance learning (MIL) [17]. In the weakly supervised multiple-instance-learning approach, each slide is considered as a bag of smaller tiles (instances) whose respective individual labels are unknown. To make a bag-level prediction, image tiles are embedded in a low-dimensional vector space, and the embeddings of individual tiles are aggregated to obtain representation of the entire image. This representation is used as input of a bag-level classifier.

For the aggregation of the tile embeddings, we used the attention-based operator proposed by Ilse et al. [18]. It consists of a simple feed-forward network that predicts an attention score for each of the embeddings. These scores indicate how relevant each tile is for the classification outcome, and are used to calculate a weighted sum of the tile representations as the aggregation operation. Weights of a bag sum to one, this way the bag representation is invariant to bag size. Finally, the bag vector representation is used as the input of a feed-forward neural network to perform the final classification.

In this approach, non-overlapping tiles of 224 × 224 pixels were extracted from each slide, and their embeddings were derived from a ResNet34 model. Empty tiles were discarded beforehand. As in the fully supervised approach, the MIL classifier was trained separately to predict IHC score and HER2 status.

To test the importance of image resolution in prediction we used a ResNet34 architecture [19] for prediction of IHC score and HER2 status. The network was trained as a four class IHC score classifier and separately as a binary classifier of the HER2 status. Given the large resolution of the tissue images (5468 × 5468 pixels), this approach required scaling them down by 5.34 to the size of 1024 × 1024 pixels to allow the network to train within our hardware memory limits.

We also constructed a method for predicting IHC score and HER2 status based on the staining intensity of the slides, a feature that is conventionally used by automatic IHC scoring software. This method was constructed to compare how predictive the single feature of staining intensity is compared to the higher level features learned by our CNN models. To extract the IHC staining expression from the images we used colour deconvolution [20]. From the staining channel, non-overlapping tiles of 224 × 224 pixels were extracted and the average staining intensity was calculated for each tile. The staining intensity of each slide was then calculated as the maximum of the average intensities of its tiles. The proposed slide descriptor was used as input in two logistic regression classifiers to predict IHC score and HER2 status separately. This approach can also be seen as a multiple-instance classification formulation where the feature extracted for each instance

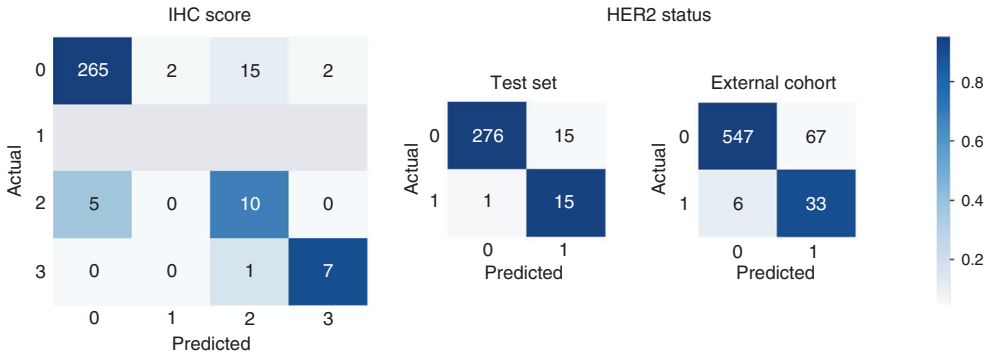



**Fig. 1 Confusion matrices of the IHC score and status prediction.** Score prediction evaluated on the test set is shown on the left, and HER2 status prediction evaluated on the test set and on the external cohort are shown in the middle and on the right, respectively.

**Table 1.** Results of the Attention-Based MIL method on the tasks of IHC score prediction and HER2 status prediction.

| Task | Balanced acc. | Precision | Recall | F1 score |
|---|---|---|---|---|
| IHC score prediction | 0.8249 | 0.9470 | 0.9185 | 0.9302 |
| HER2 status prediction | 0.9429 | 0.9705 | 0.9478 | 0.9551 |

is its average staining intensity value, and the bag is aggregated using the maximum operator.

### Network training

The dataset showed an unbalanced distribution of the IHC score (Supplemental Fig. 1) reflecting the frequency of HER2 expression in the population [21]. To obtain representative training and validation sets, we split images of each IHC score in 80-20 proportions. For the samples with score 2, the 80-20 split was done separately for those with positive status and those with negative status. During training, we performed a weighted sampling of the images of each score such that each of the IHC scores is equally represented during training. We performed random horizontal and vertical flips as data augmentation.

We used Adam optimiser in training [22], with weight decay of $1 \times 10^{-8}$ and betas of 0.9 and 0.999. The learning rates as well as their schedulers were chosen based on a hyperparameter search. The ResNet classifiers were trained using a learning rate $1 \times 10^{-5}$, which was reduced by a factor of 0.1 if the accuracy of the validation set does not improve after 20 epochs of training. The MIL classifier was trained using a learning rate of $5 \times 10^{-9}$, decreasing it by a factor of 0.3 if the accuracy of the validation set does not improve after 40 epochs. We used a batch size of 32 in the ResNet classifier and a batch size of only one full resolution image with a bag size depending on the amount of extracted tiles in the MIL classifier.

Our study is compliant with the guidelines summarised by Kleppe et al. [15]. We perform data augmentations, our test set is disjoint in time from the train set, and we demonstrate the method's performance on an external validation set. Our primary analysis was predefined and we report balanced accuracy metrics throughout this study.

Computational work was performed on the CHEOPS high performance computer, on nodes equipped with 4 NVIDIA V100 Volta graphics processing units (GPUs). We used PyTorch (version 1.8.1) [23] for data loading, creating models, and training.

## RESULTS

### IHC score prediction

First, we implemented a multiple-instance-learning (MIL) [17] method allowing us to make the classification of the images at their highest resolution. Using this technique, the images are split into smaller tiles, encoded into their numeric embeddings and ranked using the attention mechanism as proposed by Ilse et al. [18]. The attention mechanism allows for automatic identification of areas in the image that are important for the predicted score, this way providing means to inspect and interpret the prediction outcomes of the network.

This technique has shown a balanced accuracy of 0.8249, precision of 0.9470 and recall of 0.9185 (Fig. 1: left, Table 1). Given the score imbalance and the lack of samples with an IHC score 1 in the test set, the reported performance metrics were calculated in a balanced manner as an average of the metric of each individual label weighted by their number of samples of that given label. Most notably, the outermost classes 0 and 3 were predicted with the highest accuracy while ~ 33% of score 2 images were incorrectly predicted.

We next examined whether a simpler CNN-based classification approach allows for predicting the IHC score from the TMA images. In order for these images to fit within our hardware constraints, we downsampled them by a factor of 5.34 to a size of $1024 \times 1024$ pixels. We trained classification architecture ResNet34 [19] on the rescaled dataset and analysed it on the test set of images adjusted correspondingly. This approach resulted in balanced accuracy of 0.8536, precision of 0.9544 and recall of 0.8859. The almost equal accuracy and precision of this model suggests that relatively large visual details visible at a lower resolution are sufficient for the most accurate prediction.

### HER2 status prediction

We next addressed the question whether the HER2 status can be predicted from the IHC-stained images directly, without additional ISH testing. Images in our dataset with IHC score of 0 or 1 are HER2 negative, those with a score of 3 are positive. Those with a score of 2 were additionally resolved using ISH resulting in the following positive/negative HER2 status split: 77/33% in the train set, 53/47% in the test set. Out of 15 IHC score 2 images in the test set, there were eight HER2 positive and seven HER2 negative. The train-validation split was done in such a way that all the score and status combinations are distributed equally in both sets.

The MIL classifier resulted in performance with balanced accuracy of 0.9429, precision of 0.9705 and recall of 0.9478 (Fig. 1 and Table 1). As in the IHC score prediction task, the results were calculated as a weighted average of the individual metrics for class 0 (HER2 negative) and class 1 (HER2 positive) to take account of the class imbalance. Within both the HER2-negative and HER2-positive classes, less than 7% of images were misclassified resulting in balanced precision and recall >0.94. To better understand the errors of the model, we additionally inspected the HER2 status prediction accuracy within images of different IHC

scores (Table 2). With ~27% false-positive and ~7% false-negative predictions, the highest error rate occurred in images with the IHC score of 2. The higher proportion of false positives among the score 2 images could be due to the underrepresentation of samples with this IHC score and negative HER2 status in the training set in the score 2 images. In images with IHC scores 0 and 3, the prediction error was below 4%. The difference in performance between the 4-class and the binary classifiers suggests that the inter-score differences are more subtle than the ones differentiating the two HER2 statuses.

### Performance on external cohort

Even if independently, our train and test datasets were collected and prepared within one hospital. To verify how the performance of our model is dependent on the aspects related to the data preparation, we evaluated our models on an independent cohort from a different clinical centre [16]. In particular, we aimed to investigate whether HER2 status prediction is indeed possible using IHC-stained images only. The external cohort included 653 tissue samples belonging to 297 patients with the following IHC score distribution: 416/186/14/37 samples of scores 0/1/2/3 respectively. Out of the score 2 samples, 12 showed a negative HER2 status and 2 samples showed positive HER2 status.

Given the different colour distribution and potential staining quality deterioration due to the sample age, we applied a preprocessing step to these images. We used Macenko's method for stain estimation [24] together with colour deconvolution/convolution [20] to match the staining to our in-house dataset. The MIL classifier yielded a balanced accuracy of 0.8688, precision of 0.9490 and recall of 0.8908 (Fig. 1). These results support the applicability of our approach in an important clinical context where the distinction of HER2 status is key for further treatment.

**Table 2.** Cross-tabulation of true IHC score and predicted HER2 status of the test dataset. '2–' and '2+' scores stand for IHC score 2 and HER2-negative and -positive status, respectively.

| | True IHC score | | | |
|---|---|---|---|---|
| **Predicted HER2 status** | **0** | **2-** | **2+** | **3** |
| Negative | 273 | 3 | 1 | 0 |
| Positive | 11 | 4 | 7 | 8 |

### Insights into the learning process of the MIL classifier

The ResNet and the MIL classifiers achieved almost identical accuracy on our in-house test set in both the IHC score and the HER2 status prediction. However, the advantage of the more compute-intensive weakly supervised MIL approach is the possibility to inspect the visual features that the network utilises in the classification process. The embeddings and attention scores assigned to individual 224 × 224 pixel tiles can provide insights into the key visual features used by the MIL approach in the classification.

First, we examined via t-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction method [25] the embeddings of the image tiles in the test set generated by the IHC score prediction network (Fig. 2). In this visualisation, spatial proximity of tiles reflects the similarity of their embeddings. Although the network was trained on the IHC score, it also correctly separates the HER2 status of the parent TMA image. HER2-negative tiles with a

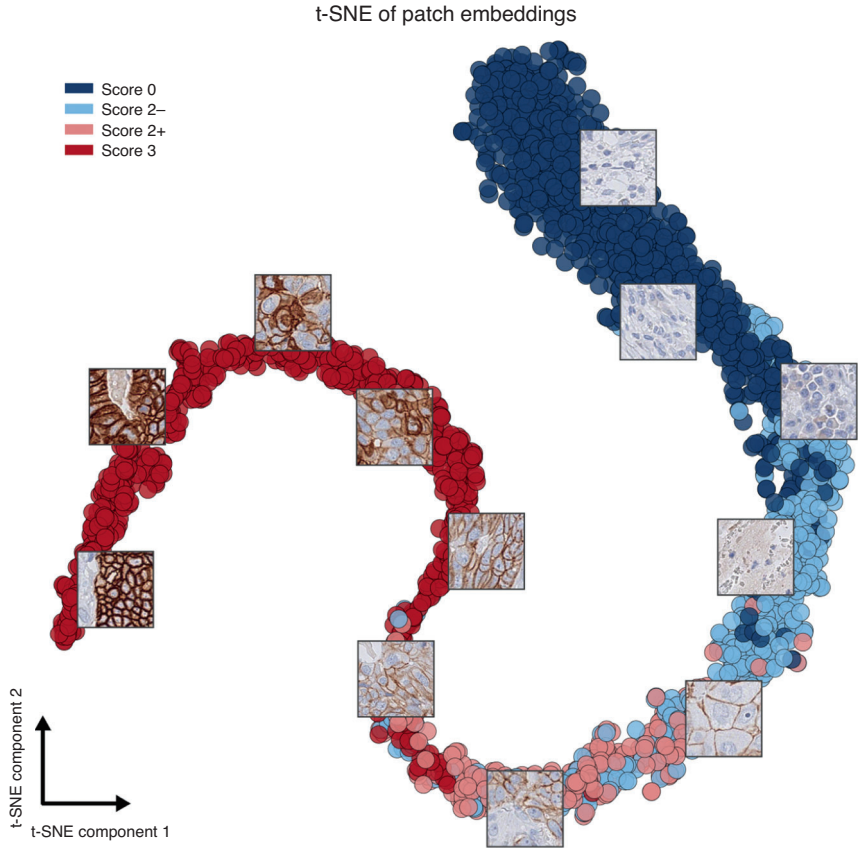



**Fig. 2 t-SNE visualisation of tile embeddings produced by the IHC score MIL classifier on the test set images, with the vectors coloured according to the score of their respective slides.** Visual similarity of the tiles is reflected in their neural network-derived representations and the embeddings of similar tiles are close in the learned vector space. Coincidentally, there are no TMA images with a score of 1 in the test set because the test set consisted of the consecutive tumour cases that followed the training set cases.

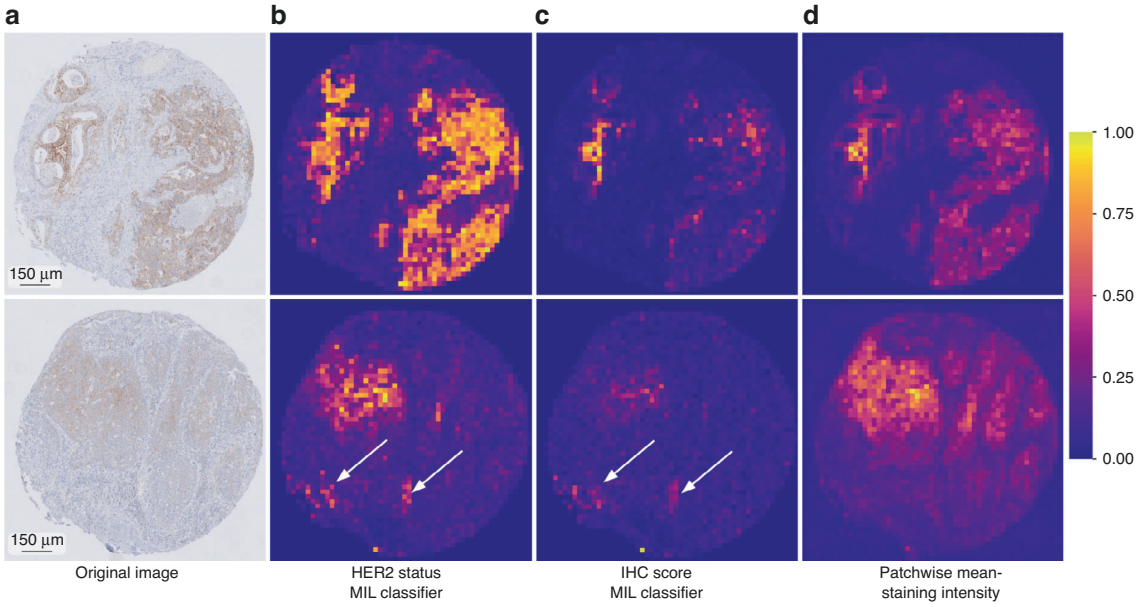



**Fig. 3 Heatmap visualisations of the attention value and mean-staining intensity in tiles within the tissue image.** The values are normalised to [0, 1]. **a** Slides with IHC score 2 and negative HER2 status. **b** Attention score heatmap of HER2 status MIL classifier. **c** Attention score heatmap of IHC score MIL classifier. **d** Patchwise mean-staining intensity heatmap. White arrows point to locations where the attention values do not match staining intensity.

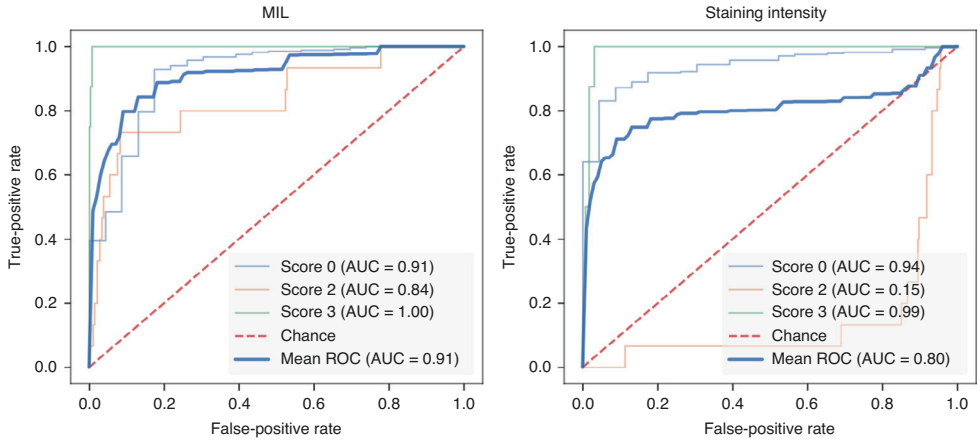



**Fig. 4 Per-class ROC curves for the IHC score classifiers, calculated in a "one-vs.-all" fashion of the MIL (left panel) and staining intensity-based classifier (right panel).** While both models' performance is similar for images of score 0 and 3, images of score 2 are not possible to correctly recognise based on staining intensity only.

score of 2 (2–) group together with score 0 tiles, and HER2-positive tiles with a score of 2 (2+) group together with score 3 tiles.

Additionally, neighbouring tiles in the t-SNE projection show visual similarity. Most strikingly, tiles grouped together show a similar staining intensity and this intensity gradually changes along the 2D projection of the embeddings. Staining intensity is, however, not the only visual feature determinant of the HER2 scoring, which also takes additional morphological features into account (potentially such as those listed in Supplemental Table 1). We expect these morphological features to also be encoded in the learned vector space.

Next, we inspected the attention values of the MIL classifier and their distribution within the tissue slides. The attention value reflects the importance of a given image tile for the final prediction score and this way provides information on spatial distribution of the visual features in the tissue that the network is exploiting in the prediction. Since the IHC staining is insufficient to resolve the HER2 status if the tissue IHC score is 2, we inspected which visual features are exploited by the network in resolving the HER2 status of the score 2 tissue slides (Fig. 3). Strikingly, the attention of the MIL classifier for the HER2 status focuses on areas of high staining intensity and corresponds to the mean intensity of the tiles at first sight.

Given the relationship of the embeddings as well as attention value to the staining intensity, we tested the accuracy of a predictive model based on the staining intensity only. Similar to the tiling approach of the MIL classifier, we split the tissue slides in $224 \times 224$ pixel tiles and averaged the staining intensity in each of the tiles. We, then, used the maximum of the average intensities across the tiles of an image as the quantitative descriptor of the entire image. We trained two logistic regression models to predict IHC score and HER2 status, respectively. The stain intensity-based model showed a balanced accuracy of 0.6876 in the prediction of the IHC score, markedly lower compared to the MIL classifier with a balanced accuracy of 0.8249. The major difference in performance between these models is in images with an IHC score of 2 (Fig. 4). In the task of predicting the HER2 status, the balanced

accuracy of the staining intensity-based model reached 0.8457 compared to 0.9429 of the MIL classifier.

These results suggest that not only the staining intensity but also additional morphological features are considered by the deep-learning models in the classification. These features are particularly important for correct recognition of images belonging to the intermediate IHC score 2. We indicate examples of such features in Fig. 3 and Supplemental Fig. 2. Even though attention value and staining intensity largely match, the heatmaps in Fig. 3 demonstrate prominent exceptions where features of high attention do not show high staining intensity.

### Comparison to existing classifiers

Several computational toolboxes currently allow for training predictive models on whole slide images (WSIs) stained using hematoxylin and eosin (H&E) [26–29]. We compared the results of our approach against CLAM [26], a publicly available pipeline for WSI classification. This pipeline extends the attention-based deep MIL proposed in [18] by including a clustering performed on the embedding space during training, which improves prediction. Similar to our approach, CLAM performs weighted sampling of images to overcome the class imbalance bias. Training and testing CLAM on the same data as our method resulted in balanced accuracy of 0.7166 (precision of 0.9479, recall of 0.7394) in the score prediction task and balanced accuracy of 0.8997 (precision of 0.9611, recall of 0.9218) in the status prediction task, markedly lower compared to our approach.

## DISCUSSION

Automated and accurate image-based diagnostics help to accelerate medical treatment and decrease the work burden of the medical personnel. Here, we demonstrate that deep-learning-based prediction of the IHC score (0–3) and the HER2 status (negative or positive) is generally possible with a balanced accuracy of ~0.85 and ~0.94, respectively. Among the scores, IHC score 2 images show the highest proportion of misclassified samples. These score 2 images cannot be unequivocally classified regarding their HER2 status by the pathologists and need further ISH-based evaluation. While it is considered that it is not possible to resolve the HER2 status based on the IHC staining of the IHC score 2 images, our models correctly predict the HER2 status of 73% of these images in our test dataset. Notably, score 2 samples are strongly underrepresented in our datasets. We expect that with more training samples of the underrepresented scores this prediction accuracy will improve.

Several computational toolboxes currently allow for training predictive models on WSIs. These multipurpose pipelines for digital pathology are crucial to the research community because they produce good results, allow for quick insights in the data with an enormous ease of use. Our comparison with an existing, publicly available WSI classification toolbox CLAM [26], suggests however that problem-tailored approaches such as ours offer refined control over parameterisation and data formatting, which allows to achieve higher accuracy and computational efficiency. Dedicated, problem-specific computational solutions might also be easier to further develop into clinical tools.

One of our key findings is that not only staining intensity—conventionally used in automated prediction tools—but also additional morphological properties are taken into account by the neural networks in the classification. We identified multiple images in which the attention maps of the MIL classifier do not match the staining intensity (Fig. 3). Additionally, prediction based on the intensity yields markedly lower accuracy suggesting that the CNN uses morphological features of the image beyond mere staining intensity. This additional information is key for the CNN to correctly predict the equivocal cases with HER2 score 2. Identification of the specific morphological signatures of HER2 not captured by the staining will require pathologists' as well as computational analysis of the high-attention and low stain intensity regions (Supplemental Fig. 2).

Neural networks for quantification of tumour morphology, especially in the H&E stainings, emerge as a novel approach for detecting tumour features invisible to the human eye, such as those corresponding to DNA mutations. Kather et al. predict microsatellite instability in gastrointestinal tumours directly from H&E stainings [30]. Couture et al. predict various breast cancer biomarkers, including the oestrogen receptor status, with an accuracy > 0.75 [31]. The authors suggest the presence of morphological features indicative of the underlying tumour biology in H&E images accessible to deep-learning methods. Lu et al. predict the HER2 status directly from H&E WSIs in breast cancer using a graph representation of the cellular spatial relationship [32] yielding an area under the receiver operator curve (AUROC) of 0.75 on an independent test set.

While inferring information imperceptible to the human eye from H&E stained tumour slides is a powerful approach pushing the boundaries of digital pathology, we use IHC-stained images in our study. Compared to H&E images, IHC stainings directly visualise the molecular HER2 expression and thus present more specific and interpretable data for pathologists. Our approach explores this information to an extent beyond human perception and staining intensity producing an AUROC curve of 0.91 (see Fig. 4). While leaving a clinical decision up to an automated method is not practiced due to its associated ethical questions, our IHC-based MIL approach could readily be used to assist pathologists. The attention maps could point clinicians to the relevant regions in the IHC images and thus save time and manual workload of clinicians.

In this study, our data is in the form of TMA, our approach is however readily applicable to WSIs and expandable to different file formats. Processing optimisations, such as precalculating tile embeddings prior to inference, might be needed if the volume of WSIs exceed the hardware memory limitations. Our results on the external test set suggest that with appropriate image normalisation our model can generalise to other datasets.

Unexpectedly, the classifiers based on low- ($1024 \times 1024$ pixel) and high- ($5468 \times 5468$ pixel) resolution images achieve matched accuracy. Potentially, the lower resolution used in this study is sufficient to encode the key morphological features of the images. This resolution was the highest that still allowed for training ResNet within our hardware memory. Notably decreasing the size of the images further to $512 \times 512$ pixel size resulted in the decrease of the model balanced accuracy to 0.8200 for the prediction of IHC score. Unlike in this study, WSIs instead of TMAs are used in the diagnostic pathological assessment. The WSI size is several orders of magnitude larger than the images in our dataset, which does not allow for using simple classification architectures such as ResNet and MIL approaches are typically used instead. Our results suggest however that reducing image resolution even 5-fold does not affect the deep-learning model performance, which could accelerate model training and reduce computational costs of models built on WSIs without compromising their accuracy.

Given the class imbalance of our datasets, we report the balanced accuracy and weighted recall, precision and F1 metrics, as the unbalanced and unweighted metrics may be misleading in describing performance of the models. As an example, if unbalanced, the accuracy score of an IHC score classifier that always predicts score 0 would be 0.92 in our dataset, and an analogous HER2 status classifier would achieve accuracy of 0.94. The unbalanced precision (and subsequently, F1) of our HER2 status classifiers would be similarly inaccurate. If we take, for example, the MIL HER2 status classifier, its unbalanced precision score is 0.51, while its false-positive rate is only 0.04. For these reasons we calculate our accuracy metrics in a balanced manner.

We propose that artificial intelligence-based HER2 status evaluation represents a valuable tool to assist clinicians. In particular, the attention map generated by the MIL classifier can aid the pathologists in their daily work by indicating the image areas of high information content for the evaluation. This approach could facilitate and speed up the manual analysis of large tissue images. The IHC score determination network can easily be transferred to any IHC staining other than HER2, further paving the way for digital pathology. We additionally demonstrate the capacity of our method to perform on samples from external clinical centres with similar prediction accuracy. We expect the power and generalisability of our deep-learning model to increase with larger, multi-centre datasets.

Finally, the high performance of our models in predicting the HER2 status of score 2 samples for which the status is considered as unresolvable based on the IHC staining, suggests that there exist visual features predictive of the HER2 status in these images. While identification of these features would require more IHC score 2 image data than available in our dataset, we expect that further deployment of the MIL models might lead to the discovery of novel morphological signatures improving image-based diagnostics.

## CONCLUSION

We demonstrate that it is possible to automatically predict HER2 overexpression directly from IHC-stained images of gastroesophageal cancer tissue, an important diagnostic process in the treatment of GEA patients. CNNs not only replicate the IHC scoring system used by pathologists, but can directly predict HER2 status in cases where it is considered not possible to resolve this condition by IHC staining alone.

Interestingly, staining intensity is not the only predictive feature for HER2 overexpression in the IHC images. Deep-learning algorithms can capture complex molecular features like the HER2 status from the tissue morphology. The attention map of the MIL classifier identifies key morphological features beyond staining intensity that might be important indicators to assess individual tumour biology.

We conclude that deep-learning-based image analysis represents a valuable tool both for the development of useful digital pathology applications and the discovery of visual features and patterns previously unknown to traditional pathology.

## DATA AVAILABILITY

The data that supports the findings of this study is publicly available in https://zenodo.org/record/7031868 [33].

## CODE AVAILABILITY

We provide our code with explanatory notebooks under https://github.com/bozeklab/HER2-overexpression.

## REFERENCES

1. Dai T, Shah MA. Chemoradiation in oesophageal cancer. Best Pract Res Clin Gastroenterol. 2015;29:193–209.
2. van Hagen P, Hulshof MC, Van Lanschot JJ, Steyerberg EW, Henegouwen MV, Wijnhoven BP, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. 2012;366:2074–84.
3. Xi M, Hallemeier CL, Merrell KW, Liao Z, Murphy MA, Ho L, et al. Recurrence risk stratification after preoperative chemoradiation of esophageal adenocarcinoma. Ann Surg. 2018;268:289–95.
4. Noordman BJ, Verdam MG, Lagarde SM, Hulshof MC, Hagen PV, van Berge Henegouwen MI, et al. Effect of neoadjuvant chemoradiotherapy on health-related quality of life in esophageal or junctional cancer: results from the randomized CROSS trial. J Clin Oncol. 2018;36:268–75.
5. Shapiro J, Van Lanschot JJ, Hulshof MC, van Hagen P, van Berge Henegouwen MI, Wijnhoven BP, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol. 2015;16:1090–8.
6. Oh DY, Bang YJ. HER2-targeted therapies—a role beyond breast cancer. Nat Rev Clin Oncol. 2020;17:33–48.
7. Wagner AD, Grabsch HI, Mauer M, Marreaud S, Caballero C, Thuss-Patience P, et al. EORTC-1203-GITCG-the "INNOVATION"-trial: Effect of chemotherapy alone versus chemotherapy plus trastuzumab, versus chemotherapy plus trastuzumab plus pertuzumab, in the perioperative treatment of HER2 positive, gastric and gastroesophageal junction adenocarcinoma on pathologic response rate: a randomized phase II-intergroup trial of the EORTC-Gastrointestinal Tract Cancer Group, Korean Cancer Study Group and Dutch Upper GI-Cancer group. BMC Cancer. 2019;19:1–9.
8. Nie J, Lin B, Zhou M, Wu L, Zheng T. Role of ferroptosis in hepatocellular carcinoma. J Cancer Res Clin Oncol. 2018;144:2329–37.
9. Qaiser T, Mukherjee A, Reddy Pb C, Munugoti SD, Tallam V, Pitkäaho T, et al. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology. 2018;72:227–38.
10. Tewary S, Mukhopadhyay S. HER2 molecular marker scoring using transfer learning and decision level fusion. J Digit Imaging. 2021;34:667–77.
11. Han Z, Lan J, Wang T, Hu Z, Huang Y, Deng Y, et al. A deep learning quantification algorithm for HER2 scoring of gastric cancer. Front Neurosci. 2022;16:877229.
12. Plum PS, Gebauer F, Krämer M, Alakus H, Berlth F, Chon SH, et al. HER2/neu (ERBB2) expression and gene amplification correlates with better survival in esophageal adenocarcinoma. BMC Cancer. 2019;19:1–9.
13. Lordick F, Al-Batran SE, Dietel M, Gaiser T, Hofheinz RD, Kirchner T, et al. HER2 testing in gastric cancer: results of a German expert meeting. J cancer Res Clin Oncol. 2017;143:835–41.
14. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7:1–7.
15. Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer. 2021;21:199–211.
16. Langer R, Rauser S, Feith M, Nährig JM, Feuchtinger A, Friess H, et al. Assessment of ErbB2 (Her2) in oesophageal adenocarcinomas: summary of a revised immunohistochemical evaluation system, bright field double in situ hybridisation and fluorescence in situ hybridisation. Mod Pathol. 2011;24:908–16.
17. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89:31–71.
18. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: International conference on machine learning. Jul 3. PMLR; 2018. pp. 2127–36.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE:2016. pp. 770–78.
20. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001;23:291–9.
21. Koopman T, Smits MM, Louwen M, Hage M, Boot H, Imholz AL. HER2 positivity in gastric and esophageal adenocarcinoma: clinicopathological analysis and comparison. J Cancer Res Clin Oncol. 2015;141:1343–51.
22. Kingma DP, Ba J. Adam: a method for stochastic optimization. https://arxiv.org/abs/1412.6980. 2014.
23. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. https://doi.org/10.48550/arXiv.1912.01703. 2019;32.
24. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. IEEE:2009. pp. 1107–10.
25. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
26. Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5:555–70.
27. van Treeck M, Cifci D, Laleh NG, Saldanha OL, Loeffler CM, Hewitt KJ, et al. DeepMed: a unified, modular pipeline for end-to-end deep learning in computational pathology. https://doi.org/10.1101/2021.12.19.473344. 2021.
28. Dolezal J, Kochanny S, Howard F, Slideflow: a unified deep learning pipeline for digital histology (1.1.0). Zenodo. https://doi.org/10.5281/zenodo.6465196. 2022.
29. Pocock J, Graham S, Vu QD, Jahanifar M, Deshpande S, Hadjigeorghiou G, et al. TIAToolbox: an end-to-end toolbox for advanced tissue image analytics. Commun Med (Lond). 2022;2:120.

30. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25:1054–6.
31. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer. 2018;4:30.
32. Lu W, Toss M, Dawood M, Rakha E, Rajpoot N, Minhas F. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. Med Image Anal. 2022;80:102486.
33. Pisula JI, Datta RR, Boerner-Valdez L, Jung JO, Plum P, Loeser H, et al. HER2 overexpression in gastroesophageal adenocarcinoma from immunohistochemstry imaging (0.1). Zenodo. https://doi.org/10.5281/zenodo.7031868. 2022.

## ACKNOWLEDGEMENTS

Both KB and JIP were hosted by the Centre for Molecular Medicine Cologne throughout this research. KB and JIP were supported by the BMBF programme Junior Group Consortia in Systems Medicine (01ZX1917B) and BMBF programme for Female Junior Researchers in Artificial Intelligence (01IS20054). This study was uploaded to bioRxiv as a preprint.

## AUTHOR CONTRIBUTIONS

JIP performed data analysis, paper writing and software development. RRD, LBV, JJ, PP, PL, MM, DPS and KL did the data acquisition. JRA contributed to software development. HL, FG, AQ performed data acquisition and analysis. CJB and KL revised the paper. AW provided the external cohort. FCP and KB designed the study and performed data analysis and writing of the paper.

## FUNDING

KB was funded by the German Ministry of Education and Research (BMBF) grant FKZ: 01ZX1917B, JIP was funded by the BMBF project FKZ: 01IS20054. Open Access funding enabled and organized by Projekt DEAL.

## COMPETING INTERESTS

The authors declare no competing interests.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was performed in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments. The present study was approved by the ethics committee of the University of Cologne (reference no. 13-091). Written informed consent was obtained from all patients.

## CONSENT FOR PUBLICATION

Not applicable.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41416-023-02143-y.

**Correspondence** and requests for materials should be addressed to Katarzyna Bozek.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 3

# Neural networks predict the pathological response to neoadjuvant radiochemotherapy in esophageal cancer from primary biopsies

---

This chapter continues the case study of EA, focusing more specifically on patients with adenocarcinoma of the gastroesophageal junction (AGEJ). It addresses the more challenging problem of predicting therapy response directly from WSIs of pre-therapy biopsies. AGEJ patients are commonly treated with perioperative chemotherapy or neoadjuvant radiochemotherapy, yet the response rates for both therapies remain moderate. Here, we investigate the feasibility of reliably identifying patients who are likely to respond positively to neoadjuvant radiotherapy by classifying biopsy slides taken at the very beginning of their medical care, thus enabling the selection of personalized treatment. In contrast to the previous chapter, this case involves no known biomarker that could serve as a target, and the task of predicting therapy response is not performed through visual inspection by histopathologists. Consequently, the biopsy slides are stained with H&E, revealing only the general microscopic anatomy of the tissue, with no specific visual cues to guide the learning process.

The methodology in this work builds upon the one presented in the previous chapter. The attention-based classifier is replaced by a transformer model, and the saliency maps are generated using an XAI algorithm known as *Integrated Gradients*, which is applicable to any gradient-based architecture. Additionally, a further analysis step is introduced, involving feature engineering and statistical analysis of the cell nuclei present in the image patches identified as relevant for the classification decision by the XAI algorithm.

At the time of writing, this chapter corresponds to the first draft of a scientific paper in preparation.

**Neural networks predict the pathological response to neoadjuvant radiochemotherapy in esophageal cancer from primary biopsies**

**Juan I. Pisula, Jin-On Jung, Yuri Tolkach, Su Ir Lyu, Alexander Quaas, Felix C. Popp, Katarzyna Bozek.**

**Abstract**

The current treatment for locally advanced adenocarcinoma of the gastroesophageal junction (GEJ) typically involves multimodal approaches, with perioperative chemotherapy or neoadjuvant radiochemotherapy demonstrating improved outcomes in clinical trials. However, only about half of the patients respond to neoadjuvant therapy. Personalized treatment strategies could improve response rates and survival outcomes if the likelihood of response could be accurately predicted.

This study explores the use of artificial intelligence (AI) to advance personalized therapy selection for GEJ cancer patients. Specifically, we hypothesize that neural networks can predict individual responses to preoperative radiochemotherapy administered according to the CROSS protocol. Using a deep learning approach, we predict therapy response to CROSS therapy—assessed by the Becker tumor regression grade—using H&E-stained tissue slides from primary GEJ tumor biopsies. Our model achieves 82% accuracy on an independent test set. Through slide segmentation and interpretability techniques, we identify key features that differentiate CROSS responders from non-responders.

By identifying patients likely to achieve pathological complete or major response, which correlates with improved survival, this AI-driven method has the potential to guide treatment decisions for GEJ cancer patients in the future. This personalized approach could optimize outcomes by selecting patients most likely to benefit from radiochemotherapy.

## Introduction

The incidence of adenocarcinoma of the gastroesophageal junction (GEJ) is rising worldwide. Patients with locally advanced GEJ cancer typically undergo a multimodal treatment approach guided by pivotal trials demonstrating improved outcomes with neoadjuvant radiochemotherapy (preoperative radiotherapy with carboplatin plus paclitaxel - CROSS) or perioperative chemotherapy with fluorouracil, leucovorin, oxaliplatin, and docetaxel (FLOT). Clinical trials such as Neo-AEGIS and ESOPEC have been conducted to compare these regimens, but they did not employ the latest therapy regimes. In the Neo-AEGIS trial demonstrating no difference between neoadjuvant radiochemotherapy and chemotherapy, most patients received chemotherapy with inferior efficacy compared to FLOT (EOX, ECX). This limitation hampers the practical application of the trial's findings in clinical decision-making. The ESOPEC study demonstrated an advantage for perioperative chemotherapy (FLOT); however, patients who received radiochemotherapy did not receive subsequent adjuvant immunotherapy, which is standard today. Additionally, only about 50% of patients responded to perioperative chemotherapy, and 56% responded to radiochemotherapy. According to the ESOPEC study, all patients would receive FLOT chemotherapy. However, since the overall response rates for both FLOT and CROSS therapies are moderate, CROSS could be a personalized treatment choice for patients whose response to CROSS can be reliably predicted.

Combination of deep learning methods and diagnostic hematoxylin and eosin (H&E)-stained histopathology images represents a promising way to extract clinically relevant insights from tumor tissue specimens. These insights include not only tumor presence and its type, which is a typical use of the images in histopathology-based diagnostics, but also molecular phenotypes for which there are no known visual biomarkers.[1,2] Similar to the detailed molecular phenotypes of cancer, there are no known visual features in the H&E slides predictive of patient outcomes. Nevertheless, the combination of histopathological images with clinical information was recently shown to allow the prediction of the survival of gastric cancer patients.[3] These studies effectively broaden the applications of H&E-stained tissue slides from their conventional role in tumor diagnosis and subtyping to discovery of novel morphological biomarkers for predicting molecular alterations and patient outcomes.

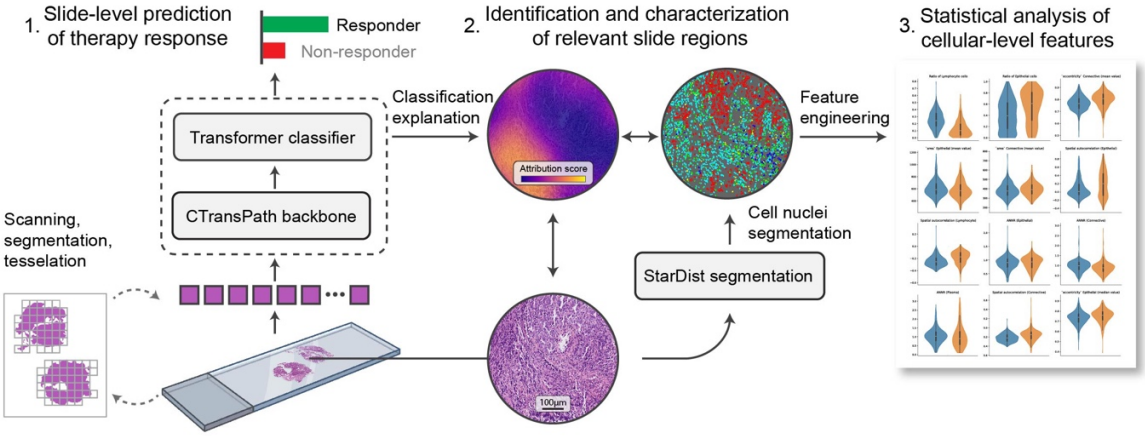



**Figure 1.** Our WSI-based response prediction workflow. In a first step, we train a transformer CROSS response prediction model. Next, we identify relevant slide regions using a post-hoc gradient-based explainability algorithm, and a nuclei segmentation model is applied to characterise their cellular compositions. We analyze several engineered features on these regions to understand the cellular factors that are associated with therapy response.

Here, we present our approach for predicting esophageal adenocarcinoma patients' response to preoperative radiotherapy based on diagnostic slides of primary tumor samples (**Fig. 1**). We measure the treatment response with the Becker tumor regression grade (TRG) and define responders as patients with pathological complete response and major response (<10% vital tumor cells). Our model predicts responders with balanced accuracy of 80.74%. Using our interpretability pipeline which combines integrated gradients with nuclei segmentation and classification, we find that image regions predictive of non-response contain more heterogenous cell populations, as well as slightly different nuclear morphology. Identification of patients who do not benefit from treatment is important to offer them alternative therapies and for establishing a personalized approach to treatment of gastroesophageal tumors.

## Methods

### *Dataset, Tumor sample, and image preparation*

We collected primary biopsies from gastroesophageal junction (GEJ) cancer patients who underwent surgery at the University Hospital of Cologne. The majority of the biopsies were originally obtained externally. We requested the corresponding tissue blocks from the respective pathological institutions and performed standardized hematoxylin and eosin (H&E) staining according to standard protocols centrally at the University Hospital of Cologne. The H&E stained slides were digitized using a NanoZoomer S360 slide scanner (Hamamatsu Photonics, Japan) at 20x magnification. We used the resulting whole slide images (WSI) of 312 primary biopsies of 193 GEJ cancer patients to train the neural network. All patients underwent CROSS therapy after surgery. The surgical specimens were used to determine the pathological tumor regression grade (TRG) according to Becker based on the percentage of vital residual tumor cells (VRTCs). The VRTC number separates the patients into four TRG categories. 1a = complete pathological regression without residual tumor; 1b < 10 % residual tumor, corresponding to major pathologic regression; 2 = 10–50% residual tumor, i.e., partial pathologic regression; and 3 > 50% residual tumor cells, corresponding to no tumor regression. Complete and major pathologic response is associated with a significant improvement in overall survival compared to no response or minor pathologic changes after neoadjuvant therapy in gastroesophageal cancer. Thus, we combined grades 1a with 1b to define responders and 2 with 3 to define non-responders for the training of the neural network (NN).

Out of the 193 patients, 83 patients (114 slides) were labeled as responders, and 113 patients (198 slides) as non-responders. For training and evaluation of the predictive model, we split the data in a stratified fashion on the patient level, making 65-15-20 splits for training, validation, and testing, respectively.

This retrospective study was conducted in compliance with the ethical guidelines approved by the ethics committee of the University Hospital of Cologne, Germany.

***Classification model***

In a pre-processing stage, the WSIs were tiled into patches of 256×256 pixels at ×20 magnification. Patches without tissue were discarded, and the remaining patches were processed with a CTransPath model to compute their feature vector representations.[4] Each WSI was treated as a collection of feature vectors corresponding to its non-empty image patches. Patients with multiple slides were treated as a single set of image patches, consisting of the union of the set of patches of their different slides.

The classification was done with a transformer encoder classifier.[5] The embedding vectors from the last layer of the transformer encoder were averaged and fed to a linear layer for the final classification. We trained our model for 2 epochs with the Adam optimizer algorithm, using a learning rate of 2.5e-4, weight decay of 5e-5, and batch size of 16. These hyperparameters were selected based on the best validation set AUROC.

***Interpretability***

Beyond mere therapy response prediction, we investigated the biological features that drive our NN classifier's decision. Our process was threefold: we detected relevant image regions responsible for the model's decision; we computed handcrafted features of the cellular composition of the image regions; and we performed the data analysis itself. This approach is described in detail below.

***Input attributions***

We used Integrated Gradients (IGs) to identify regions of a whole slide image (WSI) that play a role in the classifier's prediction.[6] IGs is a deep learning explainability algorithm that attributes the prediction of an NN to its input features. We applied IGs to our GEJ progression prediction model, to assign a positive score to image patches that contribute to the prediction of the correct class, and a negative score to patches that contribute to the prediction of the opposite outcome.

***Patch description and feature engineering***

We applied a StarDist model on the image patches located within manually delimited tumor regions of the slides of our test dataset.[7] The model was fine-tuned on the Lizard dataset of colonic nuclear segmentation,[8] which proves highly transferable to our cohort. The morphological similarities between colorectal adenocarcinoma and oesophageal adenocarcinoma ensures reliable tumor cell detection, and the model effectively identifies immune and stromal cells, as their appearance remains consistent across different tumor types. Additionally, all detections were checked by experienced pathologists to ensure proper segmentation and classification of cell nuclei. The Segmentation model classifies cells into six types: six classes: epithelial cells, eosinophils, plasma cells, connective cells, neutrophils, and lymphocytes. Once the cells in a patch were identified, we computed a total of 524 features that summarize the patch into a single feature vector. These features included:

- Cell type populations and ratios.
- Descriptive statistics (mean, median, variance, skewness, kurtosis, minimum, maximum) of nuclei morphology, such as the mean nuclei eccentricity of a given cell type, or the variance of its area. These features were computed with the `skimage.measure` Python package.[9]
- Descriptive statistics of distances between cell nuclei of different types, such as e.g. the median distance between connective cells and lymphocytes.
- Average Nearest Neighbor Ratio (ANNR) and Join Count (JC) analysis for each cell type.

The ANNR and JC features were used to quantify the spatial arrangement of cells within a patch, and they capture two different aspects of it.

ANNR was used to quantify the observed pattern of distances between cell nuclei in a patch:

$$ANNR = \underline{D_O} \,/\, \underline{D_E},$$

where $\underline{D_O}$ is the observed mean distance between each cell and its closest neighbor, and $\underline{D_E}$ is the expected mean distance between each cell and its closest neighbor if the cells were placed randomly:

$$\underline{D_E} = \frac{0.5}{\sqrt{n/A}},$$

where $n$ is the number of cells in a patch and $A$ is the patch area. An ANNR<1 indicates dense cell grouping (meaning, cells in the patch are closer than a random pattern of cells), and an ANNR>1 indicates a dispersed or evenly-spread pattern of cell nuclei. We computed the ANNR for each cell type in a patch.

JC analysis provides a measure of spatial autocorrelation: it describes how the values of a variable at neighboring spatial locations are similar to each other. In our case, the variable of interest is the cell type, where a positive spatial autocorrelation means that neighboring cells belong to the same type, and a negative spatial autocorrelation that neighboring cells belong to different classes. Spatial autocorrelation is complementary to ANNR, it quantifies neighboring cell nuclei types disregarding how close or far apart they are.

Our JC analysis was computed for each cell type individually, in the following way:

1. A patch was partitioned into a Voronoi tessellation, using the nuclei centroids as seeds for the regions.
2. The regions were binary-labeled. Given a cell type, a positive label was assigned to all the cell nuclei belonging to that class, and a negative label was assigned to the remaining regions.
3. The different types of joins were then counted. Two neighboring cells make a black-black (BB) join if they both have the positive label (i.e. the cell type being currently analyzed); a black-white (BW) join is formed between two cells of opposite labels; and a white-white (WW) join happens when two cells of the negative label neighbor each other.

This procedure was done for each cell type independently, assigning the positive label (black) to the analyzed cell type and the negative label (white) to all the other cell types. Our measure of spatial autocorrelation is given by:

$$Spatial\ Autocorrelation\ = (J_{BB}\ -\ J_{BW})\ /\ J_T,$$

where $J_{BB}$ , $J_{BW}$ , and $J_T$ are the number of BB joins, the number of BW joins, and the total number of joins, respectively. This equation is positive when the majority of joins in a patch are BB joins, indicating a positive spatial autocorrelation, and is negative when the majority of joins are BW joins, indicating negative spatial autocorrelation.

***Data analysis***

We applied IGs to all the patients in the test set and described their corresponding image patches with the features explained above. We used the patches coming from manually delimited tumor regions in the test set samples in this analysis. From all test set patches, we formed two groups: a “positive group” of image patches coming from responder patients, which were detected to be predictive of this condition based on IGs; and a “negative group” of patches coming from non-responder patients, which were detected to be predictive of this condition based on IGs.

To enhance the predictive signal and avoid over-representing patients with larger tumors, we used the top 10% IGs-scored patches in a slide, and limited their number to 200 image patches per slide. We compared values of each feature individually between the two groups of patches. We guide our analysis by focusing on features whose values differed between the two groups with an Effect Size bigger than random. We used the Common Language Effect Size (CLES),[10] or probability of superiority, as it has no assumptions about the data distribution, and is straightforward to understand:

$$CLES = P(X > Y),$$

is the probability that a value sampled from group X is bigger than a value sampled from group Y. In our case, the two groups were the positive and the negative groups previously described, and we computed the CLES for each feature with brute force, by exhaustively comparing each value of one group with all the values of the same feature in the other group.

## Results

### *Classification model predicts therapy response based on primary biopsies*

We predict patient response to preoperative radiochemotherapy using pre-treatment histopathological images of primary biopsies. To assess therapy response, we used the Becker tumor regression grade (TRG) defined by vital residual tumor cells. A pathological complete response (pCR, no vital tumor cells) and major response (mPR, <10% vital tumor cells) are associated with a significant improvement in overall survival in gastroesophageal cancer. We categorized patients as therapy responders by combining those with pCR and mPR, while considering the remaining patients as non-responders. This classification resulted in the establishment of two distinct classes. We split the 312 images of 193 GEJ patients into a training set (n = 123), a validation set (n = 31), and an independent test set (n = 39). Patient-level data splitting prevented bias by ensuring that multiple slides from one patient were grouped in the same dataset.

Our classification model achieved AUROC of 0.80 on the test set (**Fig. 2**). We observed robust performance metrics when evaluating the test set. with sensitivity of 70.59%, and the specificity of 90.91%. We found the precision to be 85.71%, indicating a reliable likelihood of identifying responders. Correspondingly, the negative predictive value (NPV) was 80.00%, affirming the network's ability to exclude non-responders (see Figure 2). In summary, our network exhibits suitable sensitivity, specificity, and predictive values, supporting its effectiveness in accurately identifying responders and non-responders to CROSS treatment. The evaluation in the validation set yielded a similar performance, measuring AUROC of 0.82, sensitivity of 76.92%, specificity of 88.89%, precision of 83.33%, and NPV of 84.21%.

We presented the network with 79 images of patients who underwent FLOT treatment. The network accurately identified 20.00% of the responder cases, with a precision of 80.00%. This suggests that our model is indeed CROSS-specific and shows a strong capability to distinguish CROSS responders while making conservative decisions for FLOT response.

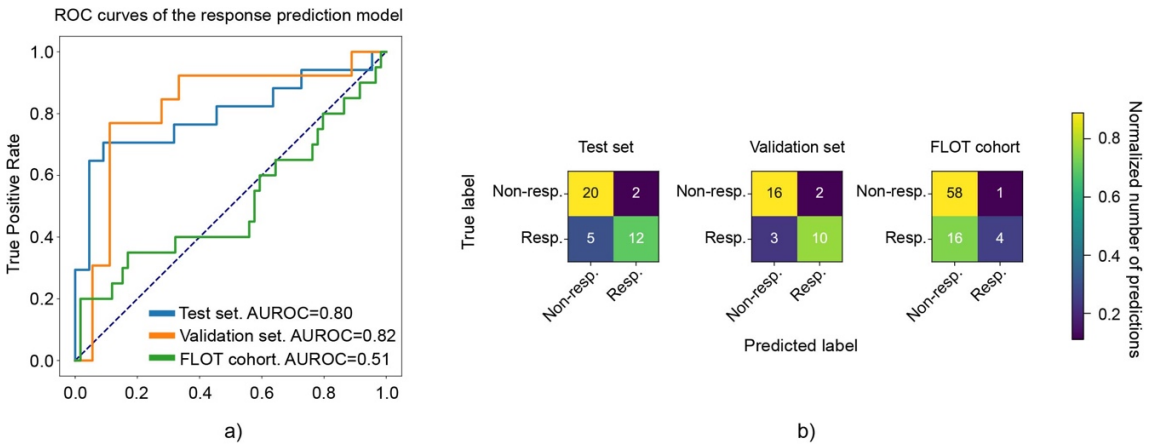



**Figure 2.** a) ROC curves of our response prediction model. b) Confusion matrices of our response prediction model.

***Interpretation of the predictive model***

In contrast to tumor detection and subtyping, there are no known visual biomarkers predictive of GEJ cancer patient response to treatment. Given the high accuracy of our predictive model we next devised an interpretability pipeline to unravel which morphological and tissue structure features are predictive of patient response to CROSS therapy. Briefly, we identified which WSI patches are most important for prediction using Integrated Gradients, we also segmented and classified all cell nuclei within tumor regions in our test set (**Fig. 3**). Using segmentation results we quantified 524 features describing cell and tissue morphology and organization and compared these features between two groups of image patches: those strongly predictive of response and of the lack of response in the two patient groups, respectively. We ranked features based on the effect size in this comparison.

This analysis revealed distinct cellular and structural differences between patches predictive of response and predictive of lack of response. We found that the most discriminant feature is the composition of cell populations. Patches predictive of response contained fewer lymphocytes and connective cells, but more tumor cells compared to those predictive of non-responders (**Fig. 4a-c**). This aspect is also reflected in the spatial autocorrelation of tumor cells, with non-responder patches having more interactions between tumor and other cell types (**Fig. 4d**). In contrast, in patches predictive of responders there is an elevated number of interactions

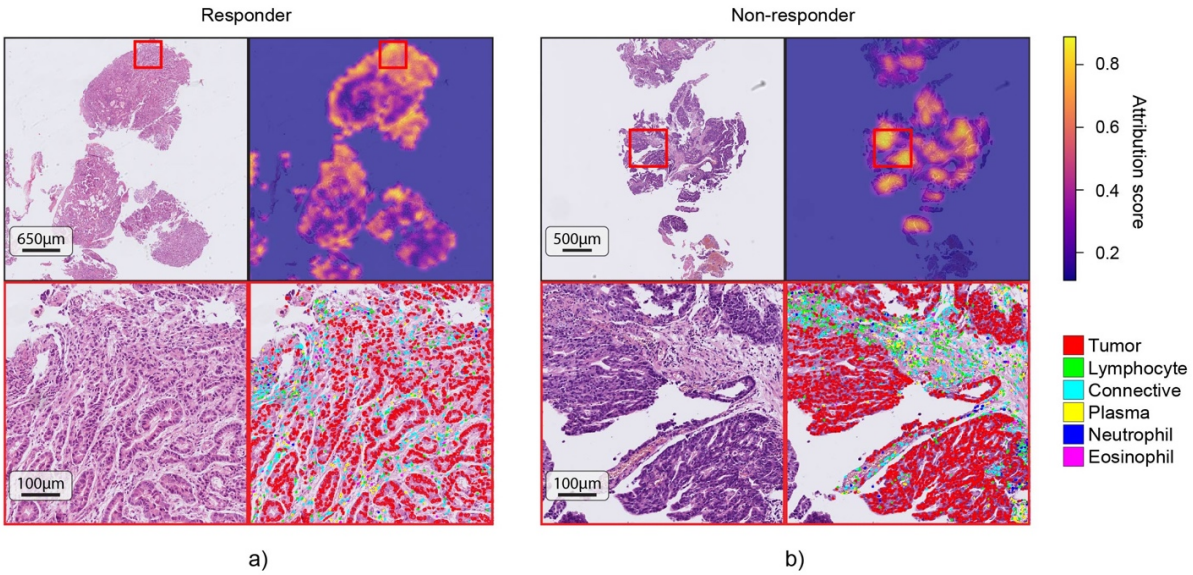



**Figure 3**. Slide, attribution heatmap, Delimited region of interest, and nuclei segmentation output. a) Responder. b) Non-responder.

between tumor cells, suggesting their spatial clustering. When inspecting the complete tumor regions of the complete dataset, we find a similar pattern in the populations of lymphocytes and tumor cells (**Suppl. Fig. 2**). Additionally, we find that connective and tumor cells are slightly more eccentric in responder-predictive patches (**Fig. 4e,f**). Representative patches of these findings are shown in **Fig. 4g,h**.

Additional violin plots of features computed on predictive patches are shown in **Suppl. Fig. 1**. Interestingly, there is no statistical difference in the tumor cell count between patches predictive or response and no-response. Connective cells are more eccentric and slightly bigger in responder-predictive patches. Distances between lymphocytes show a smaller range in responder patches. Consistently with the positive spatial autocorrelation of tumor cells in responders, lymphocytes and connective cells in non-responders have a lower spatial autocorrelation indicating higher interactions with other cell types.

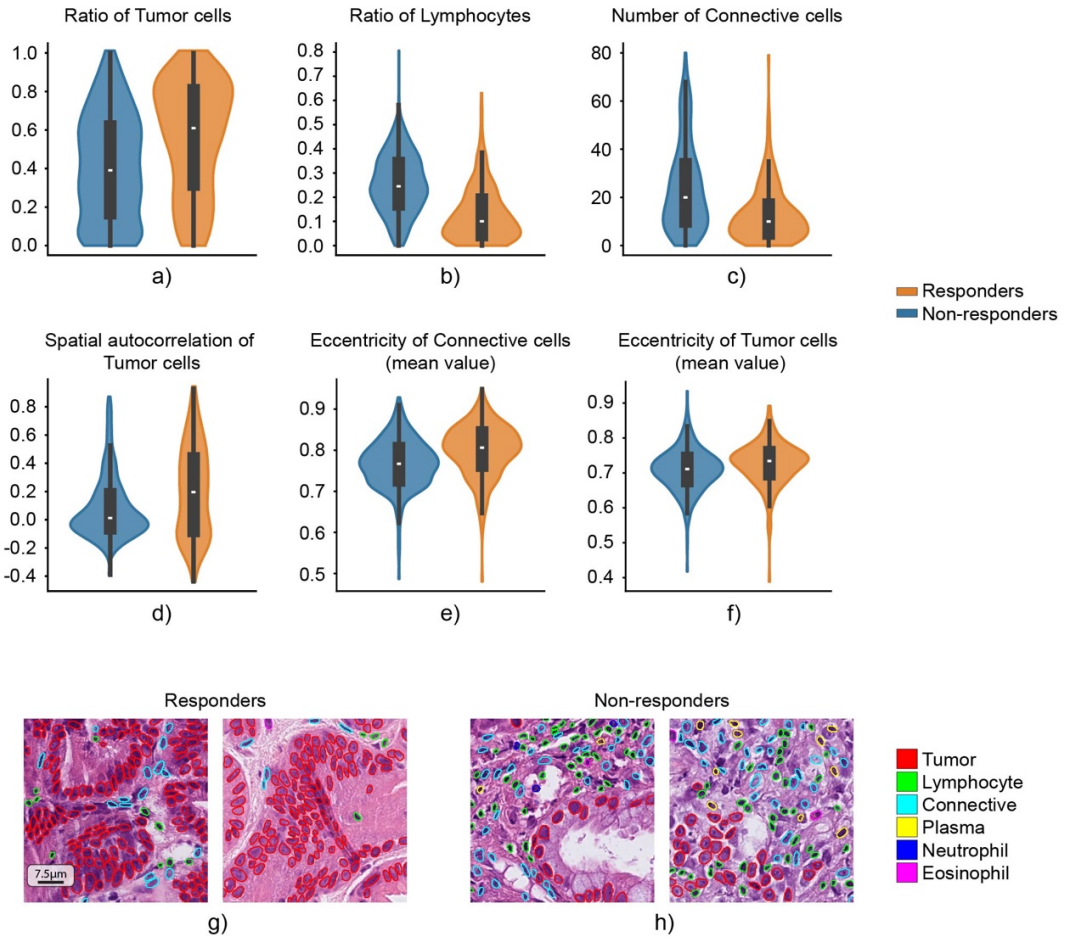



**Figure 4.** a-f) Violin plots of features which differ between the most predictive patches of responders and non-responders. g,h) Representative patches of both groups. All shown features are statistically significant according to a two-sample Mann-Whitney U-test with $p<0.0005$.

**Discussion**

While deep learning models have reached human-level performance in diagnostic pathology tasks such as tumor detection and subtyping, the problem of image-based prediction of therapy response is a remarkably more challenging task. In our study, we leverage advances in machine vision to address the critical challenge of predicting therapy response to radiochemotherapy according to the CROSS protocol for patients with GEJ adenocarcinoma using H&E-stained WSIs obtained pre-treatment during routine diagnostic workup. While for this hard-to-treat tumor, perioperative chemotherapy following the FLOT protocol is becoming the standard of care, radiochemotherapy remains a viable alternative. Given the moderate response rates to both treatments, a personalized therapy approach with the CROSS protocol could significantly benefit patients - provided the therapy response can be reliably predicted.

Here, we introduce a deep-learning approach for identifying gastroesophageal junction (GEJ) adenocarcinoma patients who respond to CROSS treatment. Our method achieves an AUROC of 0.80 on an independent test set, demonstrating high sensitivity and specificity. Notably, the model trained to predict tumor regression following CROSS treatment exhibits significantly lower performance when predicting responses to FLOT therapy. This discrepancy highlights the distinct mechanisms underlying the two treatments and underscores the need to develop dedicated prediction models tailored to FLOT therapy.

Together with the predictive model, we have developed a comprehensive and quantitative interpretability pipeline that translates model predictions into interpretable morphological and tissue organization features. We discover that tissue regions predictive of CROSS response have a more homogeneous cell population consisting mainly of tumor cells, showing a positive spatial autocorrelation and having less interactions with other cell types. On the other hand, tumors with more dispersed structures, which are intermixed with immune cells represent an environment where CROSS therapy has limited efficacy.

The size of our dataset is limited and calls for further validation with larger, multi-center cohorts. Nevertheless, our results suggest that prediction of GEJ cancer patient response to radiochemotherapy based on the primary biopsy is indeed possible. A systematic approach to

interpretation of the predictive model can point to key morphological features that distinguish the two patient groups and allow them to undergo alternative treatments.

**References**


1. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. 25, 1054–1056 (2019).
2. Tian, F. et al. Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. Nat. Med. 30, 1309–1319 (2024).
3. Muti, H. S. et al. Deep learning trained on lymph node status predicts outcome from gastric cancer histopathology: a retrospective multicentric study. Eur. J. Cancer 194, 113335 (2023).
4. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559 (2022).
5. Vaswani, A. et al. Attention is All you Need. in Advances in Neural Information Processing Systems vol. 30 (Curran Associates, Inc., 2017).
6. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. Preprint at http://arxiv.org/abs/1703.01365 (2017).
7. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell Detection with Star-Convex Polygons. in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 (eds. Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C. & Fichtinger, G.) 265–273 (Springer International Publishing, Cham, 2018). doi:10.1007/978-3-030-00934-2_30.
8. Graham, S. et al. Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. in 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) 684–693 (IEEE, Montreal, BC, Canada, 2021). doi:10.1109/ICCVW54120.2021.00082.
9. Walt, S. van der et al. scikit-image: image processing in Python. PeerJ 2, e453 (2014).
10. McGraw, K. O. & Wong, S. P. A common language effect size statistic. Psychol. Bull. 111, 361–365 (1992).

Chapter 4

# Explainable, federated deep learning model predicts disease progression risk of cutaneous squamous cell carcinoma

---

The following work builds upon the methodology presented in the previous chapter, now with the goal of predicting progression risk in patients with cutaneous squamous cell carcinoma. This disease is routinely treated with surgical excision to remove the malignant tumor, a procedure that is successful in most cases. However, a minority of patients experience local recurrence or metastasis, and current risk stratification systems cannot accurately predict these outcomes.

This paper presents two main differences compared to the previous chapter. First, the MIL model is trained using Federated Learning, a method for collaborative model training in which data remains within its originating medical center. Second, the MIL classifier is a transformer model pre-trained on a large corpora of text data to model languagea model that will be described in detail in the next chapter.

At the time of writing, this work is currently under review by a scientific journal.

# Explainable, federated deep learning model predicts disease progression risk of cutaneous squamous cell carcinoma

**Juan I. Pisula[1,2]*, Doris Helbig[3,]*, Lucas Sancéré[1,2], Oana-Diana Persa[4], Corinna Bürger[2,5,6], Anne Fröhlich[7], Carina Lorenz[2,5,6], Sandra Bingmann[8], Dennis Niebel[9], Konstantin Drexler[9], Jennifer Landsberg[7], Roman Thomas[5,10,11], Katarzyna Bozek[1,2,12]*†, Johannes Brägelmann[2,5,6]*†**

[1] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany

[2] Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany

[3] Department for Dermatology, University Hospital Cologne, Cologne, Germany

[4] Department of Dermatology, Technical University Munich, Munich, Germany

[5] University of Cologne, Faculty of Medicine and University Hospital Cologne, Department of Translational Genomics, Cologne, Germany

[6] University of Cologne, Faculty of Medicine and University Hospital Cologne, Mildred Scheel School of Oncology, Cologne, Germany

[7] Department of Dermatology and Allergology, University Hospital Bonn, Bonn, Germany.

[8] Department for Dermatology, University Hospital Cologne, Cologne, Germany

[9] Department of Dermatology, University Medical Center Regensburg, Regensburg, Germany

[10] Institute of Pathology, Medical Faculty, University Hospital Cologne, University of Cologne, Cologne, Germany

[11] DKFZ, German Cancer Research Centre, German Cancer Consortium, Heidelberg, Germany

[12] Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany

* Equal contribution

† Corresponding

**Abstract**

Predicting cancer patient disease progression is a key step towards personalized medicine and secondary prevention. The ability to predict which patients are at an elevated risk of developing local recurrences or metastases would allow for tailored surveillance of these high-risk patients as well as enhanced and timely interventions.

We developed a deep learning transformer-based approach for prediction of progression of cutaneous squamous cell carcinoma (cSCC) patients based on diagnostic histopathology slides of the tumor. Our model, trained in a federated manner on patient cohorts from three clinical centers, reached an accuracy of AUROC=0.82, surpassing the predictive power of clinico-pathological parameters used to assess progression risk. We conducted an interpretability analysis, systematically comparing a broad range of spatial and morphological features that characterize tissue regions predictive of patient progression. Our findings suggest that information located at the tumor boundaries is predictive of patient progression and that heterogeneity of tissue morphology and organization are characteristic of progressive cSCCs.

Trained in a federated fashion exclusively on standard diagnostic slides obtained during routine care of cSCC patients, our model can be deployed and expanded across other clinical centers. This approach thereby offers a potentially powerful tool for improved screening and thus better clinical management of cSCC patients.

## Introduction

Cutaneous squamous cell carcinoma (cSCC) is the second most prevalent type of non-melanoma skin cancer that is diagnosed in 1 million patients in the USA every year.[1] In the last decades, the incidence of cSCC has risen sharply and is projected to increase further.[2] Even though the majority of cSCCs can be removed by surgical excision, a relevant fraction of patients experience disease progression by local recurrence or metastases to lymph nodes or other body sites, which is associated with poor prognosis and increased risk of death.[3–6] Due to the high incidence of cSCC, this poses a significant public health concern. Reliable predictors are thus needed to decide which patients will benefit from enhanced secondary prevention e.g. by more frequent follow-up care or additional treatments such as immuno-, chemo- or radiotherapy. Current cSCC staging systems like the American Joint Committee on Cancer (AJCC), the Brigham Women's Hospital (BWH), or the National Comprehensive Cancer Network (NCCN) staging systems aim to provide guidance on risk stratification and clinical management of cSCC patients.[7,8] However, they fall short of reliably identifying patients at high risk of disease progression. Recently, multi-gene expression signatures have been used to predict metastasis risk of cSCCs.[9,10] While these signatures help to predict metastasis risk, they have not yet been used to predict local recurrences. In addition, they require measurement of gene expression from patient samples, which limits their potential for translation into clinical routine use.

In addition to clinical parameters such as immunosuppression, several pathological tumor features such as perineural involvement, tumor size, and invasion depth have been associated with increased risk of cSCC progression.[4–6] Moreover, specific histological subtypes e.g. desmoplastic cSCC have been linked to higher recurrence and/or metastasis risk.[6] Morphology in histological specimens thus holds information on progression risk, but has not yet been exploited systematically. Since deep learning has matched human experts in cancer detection and classification,[11] computational pathology methods hold promise to extract information on patient progression from histopathology image data. Building robust models that offer high predictive power across data independent of their source, requires multi-institutional data sets for model training. Obtaining such data sets poses challenges regarding data governance and raises concerns about patient privacy. Federated Learning (FL) is a strategy that limits the

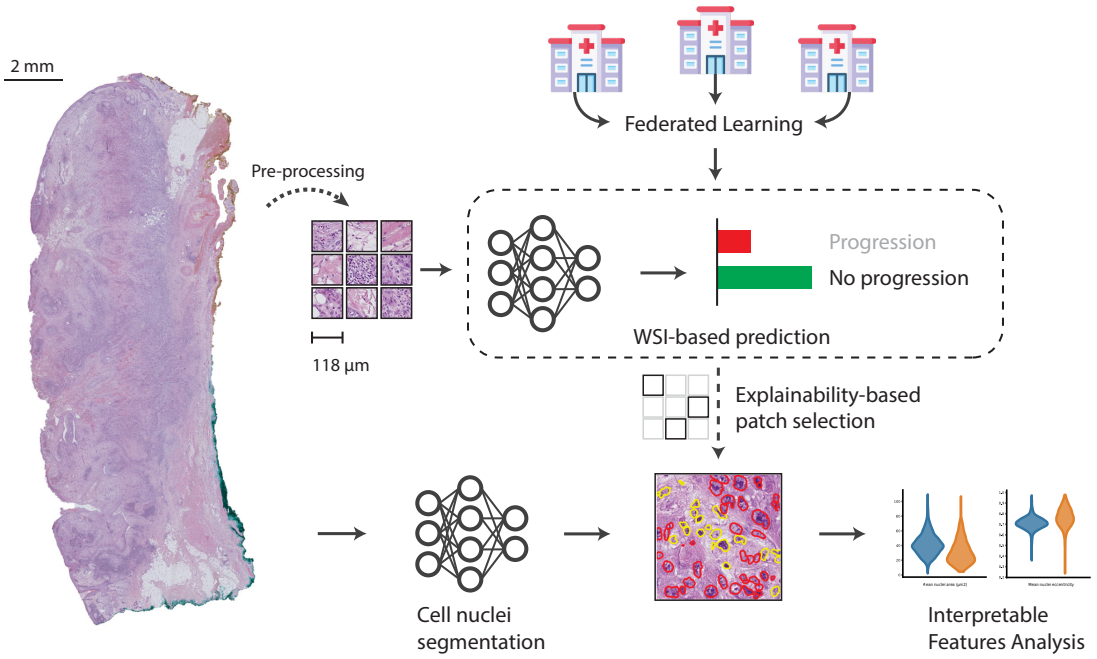



**Figure 1:** We propose a WSI-based cutaneous Squamous Cell Carcinoma (cSCC) progression prediction model, trained on data from three medical centers using Federated Learning. Beyond prediction, we investigate underlying biological features that influence our classifier. We do so by computing cellular-level features with aid of a nuclei segmentation model. We analyze these features in image regions detected as relevant for prediction outcome by Integrated Gradients, an input attribution algorithm for explainable deep neural networks.

logistic overhead and reduces privacy concerns in training a multi-center-based model.[12,13] Moreover, FL simplifies the inclusion of new patients and cohorts for further model training, which in turn facilitates model update, continuous improvement, and clinical applicability.

Here, we present a multiple instance learning transformer-based deep learning model for prediction cSCC progression risk using Hematoxylin-Eosin-(HE-) stained histopathology images acquired during routine care (**Fig. 1**).[14,15] Our model, trained in a federated manner on cohorts from three clinical centers, achieved high accuracy in predicting patients at risk of disease progression, which corresponds to significant differences in progression-free survival. We developed explainability methods on our model which provide insights into the tissue areas and cell features associated with increased progression risk. Overall, we present a powerful

approach that improves risk-stratification of cSCC patients and offers insights into the underlying cancer biology.

## Results

### Deep learning on histopathology images predicts cSCC progression risk

Currently, it is not clear if the progression risk of a cSCC can be inferred from a histopathology slide and if so, which elements of the tumor and its microenvironment are decisive of disease progression. To fill this gap, we used a multiple instance learning, transformer-based classifier for the task of progression prediction from Whole Slide Images (WSIs). We trained the model in a federated manner, leveraging data from three different medical centers (**Fig. 1**).[14,15]

Initially, we trained our model on the Cologne cohort only (n=157 patients, 214 WSIs), achieving cSCC progression status classification accuracy of 0.92 AUROC (95% CI=[0.83-1.00]) in a held-out test set from Cologne (**Fig. 2A)**. In comparison, a multivariable logistic regression model incorporating clinico-pathological parameters associated with risk of disease progression (**Suppl. Fig. 1**) achieved an AUROC of 0.64 (95% CI=[0.52-0.75]) in the same prediction task and cohort (**Fig. 2B**). To test the robustness of our deep learning model we assembled two additional cohorts from dermatology departments at the University Hospital Bonn (Bonn cohort, n=35 patients, 133 WSIs) and the Technical University Munich (Munich cohort, n=51 patients, 113 WSIs). While the model trained on the Cologne cohort performed well on the Bonn cohort (AUROC=0.90, 95% CI=[0.71-0.97]), it failed to generalize to the Munich cohort (AUROC=0.46, 95% CI=[0.30-0.63]; **Fig. 2A**). This highlights that variation induced by e.g. technical procedures or distribution shift and domain adaptation problems may hamper generalizability of models trained on a single-center cohort.

### Federated learning improves generalizability of image-based classification

To improve performance across cohorts, it is crucial to train deep learning models on large and diverse datasets. However transfer of patient data and histological slides across hospitals carries important logistic complexity and poses potential privacy threats. We therefore trained our model in an FL scheme on all three cohorts (**Fig. 1**).[12] FL overcomes the data sharing hurdles by reducing the organizational overhead of combining different patient cohorts, since patient data can remain in the respective hospital. Model training is performed locally and only model parameters are shared between the hospitals. Moreover, it enables dynamic patient enrollment and facilitates inclusion of additional centers, which in turn

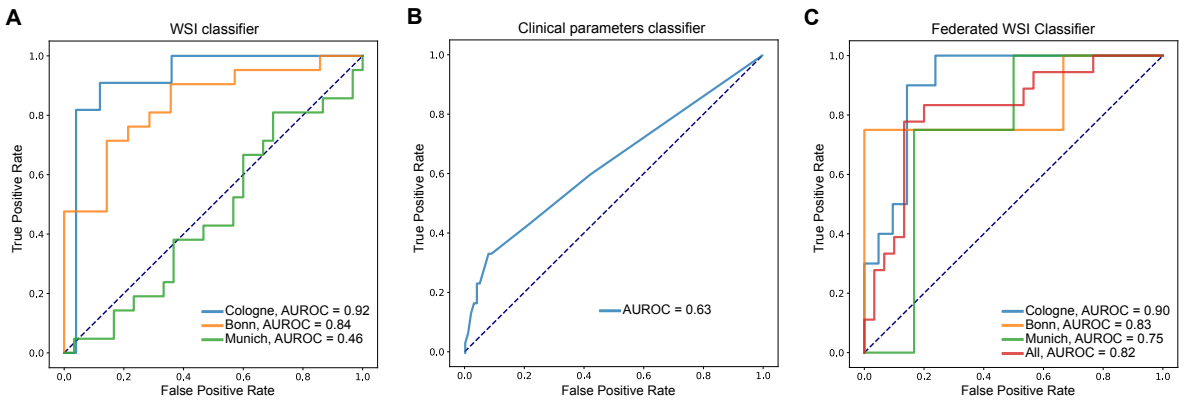



**Figure 2:** ROC curves of the classifiers. **A**: WSI-based classifier trained exclusively on the Cologne cohort and tested on Munich and Bonn cohorts (AUROC = Area under the receiver operator curve). **B**: Multivariate logistic regression model based on clinico-pathological parameters associated with progression risk in univariate analysis. Model trained and evaluated on the Cologne cohort. **C**: Federated WSI-based classifier.

increases its flexibility and the opportunities for clinical deployment. Training on the multi-institutional cohorts using the FL framework did indeed improve model performance. While AUROC on Cologne and Bonn decreased at most by 2%, performance on the Munich cohort increased by 63%, leading to prediction accuracy of AUROC=0.82 (95% CI=[0.69-0.95]) in the complete dataset (**Fig. 2C**). This highlights that prediction of disease trajectories is indeed possible for cSCC patients and can be achieved with a deep learning model trained on different cohorts in a federated manner. Such prediction opens possibilities for clinical translation of the model as a tool for the identification of patients at high recurrence risk that may benefit from increased surveillance.

**Explainability analyses highlight factors associated with cSCC progression**

In addition to stratifying patients according to their disease progression risk, we assessed which parts of the histological images are predictive of disease progression. We used Integrated Gradients (IGs) attributions to infer which areas in the WSIs are the most relevant for the prediction of the respective patient as progressor/non-progressor.[16] Additionally, we leveraged a pipeline we recently established specifically for cSCC, which performs nuclei segmentation and classification of cells into one of six cell types (granulocyte, lymphocyte, plasma, stroma, tumor, and epithelial cell).[17] We used the cell type detection and classification

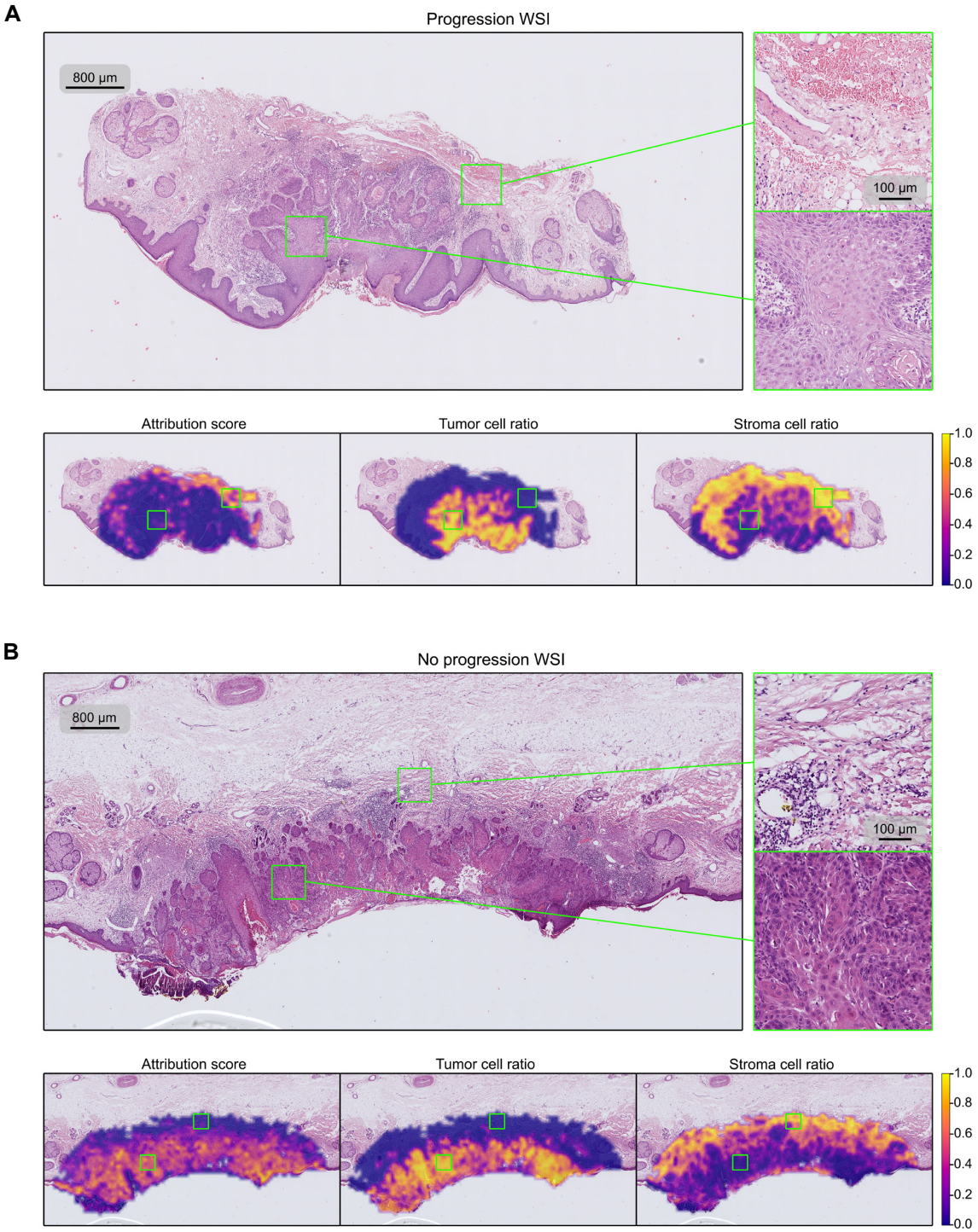



**Figure 3**: Slides and heatmaps of the patches' classifier attribution score, tumor cell ratio, and stroma cell ratio. **A:** Slide of a progression patient, showing that the WSI-based classifier assigns higher importance to the region outside the tumor area (indicated by the tumor cell ratio heatmap). **B:** Slide of a non-progression patient, where the high attribution area coincides with the tumor-cell populated areas. Colorbar indicates the slide-normalized heatmap values.

to analyze the WSI regions with the highest predicted power as attributed by IGs. In the WSI regions with high IGs attribution score we calculated various features of nuclei morphology, cell type composition and spatial distribution (**Suppl. Table 1**).

We next performed statistical analyses of these features to gain insights into the determinants of cSCC progression. Interestingly, many of the predictive tiles with the highest attribution score for disease progression were outside of the tumor region (**Fig. 3A**). In fact, attribution scores were low in areas with high tumor cell density, as determined using our cell type classification pipeline (**Fig. 3A, bottom left & middle**).[17] Instead, they were high at the tumor border and frequently in areas where the most common cell type was stroma (**Fig. 3A, bottom right, Suppl. Fig. 2**).

In contrast, for patients without disease progression, the most predictive tiles were located within the tumor and in areas with high tumor cell density. Areas outside the tumor border were, in the case of these patients, not of high value for prediction of non-progression (**Fig. 3B**). This highlights that different parts of histological sections contain information that distinguishes patients at high vs. low risk of disease progression and that such patient stratification needs to be based not only on the tumor but also its surroundings for adequate predictions.

Additionally, we systematically compared the cell-based features between the tiles that were regarded as most predictive for disease progression or non-progression according to their IGs scores. Numerous parameters with significantly different distributions between the two groups were detected, **Fig. 4** shows a subset of the tumor-cell-related features. Non-progressors e.g. showed higher values in Average Nearest Neighbor Ratio (ANNR), indicating a higher uniformity in the way tumor cells were distributed (**Fig. 4A, p<0.0001**), while progressors had more intermixing of tumor cells with other cell types, i.e. more heterogeneity in tissue composition (**Fig. 4C, p<0.0001**). Moreover, tumor cells of non-progressors showed differences in their morphology compared to progressors such as larger nucleus size (**Fig. 4B, p<0.0001)** and lower nuclear eccentricity (**Fig. 4D**, **p<0.0001**). In addition, tumors of patients that later experienced disease progression showed higher degrees of nuclear dysmorphia and pleomorphism compared to non-progressors. Tumor cells from non-progressors have larger values of morphological solidity and extent (larger median, negatively-skewed distributions,

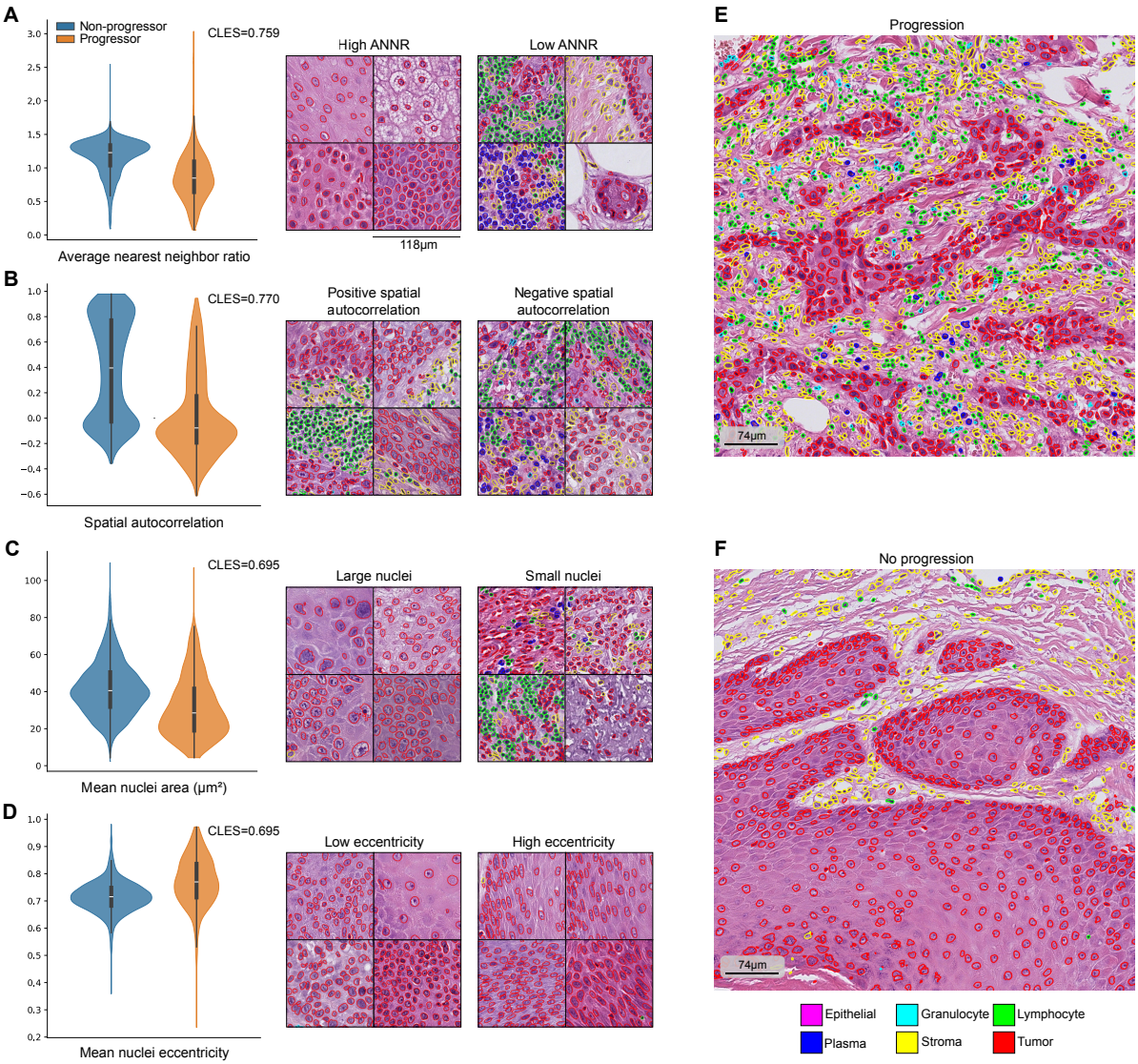



**Figure 4:** Four of the features of the tumor cells used in the analysis. **A,B,C,D** show violin plots and segmented image patches that illustrate these values. In general, progression-associated tumor cells cluster together **(A)**, interface with other cell types **(B)**, and have smaller **(C)**, eccentric nuclei **(D)**. These effects are not just local to image patches, but they occur in larger regions, as shown in **E**,**F**. The displayed CLES (Common Language Effect Size) values are indicated for the group with the largest mean. All features are significantly different in both groups, with p-values < 0.0001 using Mann-Whitney U test.

**Suppl. Fig 3A-D, Suppl. Table 1**), while morphological extent has a larger variance in tumor cells from progressors **(Suppl. Fig 3E, Suppl. Table 1)**.

We next tested whether our cell-based features are sufficient to predict the progression/non-progression of patients based on their respective image tiles using a tree-based classification algorithm XGBoost.[18] Interestingly, using the cell-based features as input resulted in high prediction accuracy (**Suppl. Fig. 4,** AUROC=0.98, 95% CI=[0.97-0.99]). This highlights that these features, which we computed using an independent pipeline, do indeed capture relevant biological parameters and variation associated with progression risk of patients. Thus the cellular and morphological features are making explicit the morphological and structural components of the tissues and cells that the deep learning model learned implicitly.

Overall, our explainability analyses indicate that tumor cell-intrinsic properties as well as composition of the microenvironment and growth patterns of the tumor are associated with the difference in prognosis and are captured by our deep learning model to accurately predict progression risk.

**Image-encoded information has higher discriminative power than clinical variables**

Several clinico-pathological parameters have been associated with increased risk of disease progression, such as immunosuppression, perineural involvement, tumor size, and invasion depth.[4,6] Similarly, desmoplastic cSCC histology has been linked to higher recurrence and/or metastasis risk.[6] We used the clinico-pathological parameters available for the Cologne cohort to test their associations with survival and to compare their predictive power with the accuracy of the deep learning model. In this experiment, we used the logit output of the deep learning model as a progression risk score. Among clinico-pathological parameters, perineural invasion and beyond subcutaneous invasion were significantly associated with shorter progression free survival in univariate analyses (**Suppl. Fig. 5A, B**). Other parameters such as thickness >6mm, ulceration, and higher grade showed trends towards shorter survival, but did not reach significance (**Suppl. Fig. 5A**). Even among high-risk patients with perineural invasion or invasion beyond subcutaneous tissue not all patients developed disease progression, i.e. recurrence or metastasis. On the other hand, among the patients with one of those risk factors our deep learning model correctly separated those who progressed from those who did not based on their predicted progression risk (**Fig. 5A**). Similarly, deep learning-based predicted risk scores were higher for patients that experienced disease progression independent of

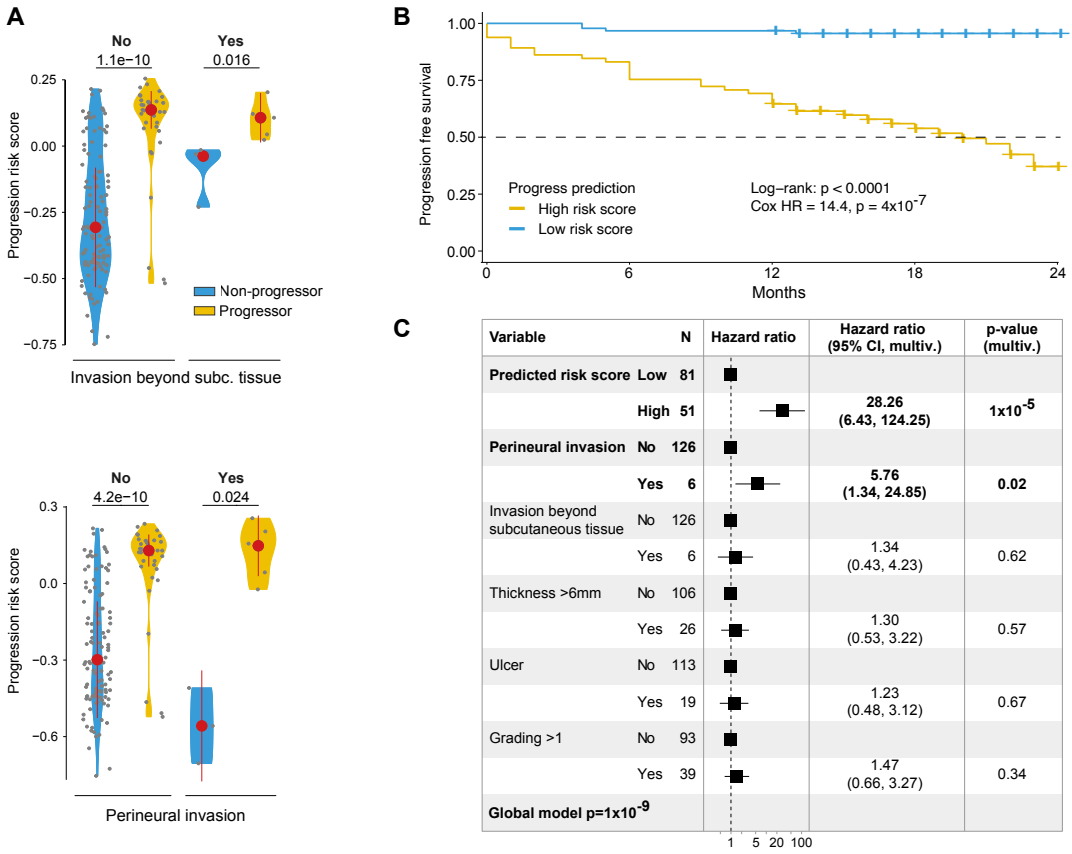



**Figure 5**: Comparison of deep learning-based classification with clinico-pathological parameters. **A:** Comparison of deep learning-based progression risk scores in Cologne patients with or without cSCC progression stratified by presence of invasion beyond subcutaneous tissue (top) or perineural invasion (bottom). Shown are median and median absolute deviation. p-values calculated by t test. **B:** Progression free survival of patients classified as high vs. low progression risk based on deep learning-based risk prediction. (Threshold determined by Youden index). Hazard ratio (HR) between groups calculated using univariate Cor regression model. **C:** Multivariable Cox regression model for n=132 Cologne patients with available data combining deep-learning based risk category with clinical parameters associated with progression free survival in univariate analyses. Shown are Hazard ratios, 95% Confidence intervals (CIs) and multivariate p-values.

tumor thickness and across histological grades (**Suppl. Fig. 5C**). The model thus allows further differentiation of patients compared to clinico-pathology-based risk factors. Without explicitly measuring these risk factors our model encodes information allowing to differentiate the two groups of patients with increased accuracy.

We next inspected the relationship of predicted risk of progression to patient survival. Using only the deep learning model's predicted progression risk score to classify patients as high- vs. low-risk, stratifies patients with short from those with long survival (median PFS 20.1 months vs. not reached in high vs. low risk, respectively; **Fig. 5B**). The risk of progression was 14 times higher for high- compared to low-risk patients (Hazard ratio 14.4, **p<0.0001**; **Fig. 5B**). Similarly, dividing patients into risk groups based on tertiles of the deep learning-based progression risk scores reaches similar performance (**Suppl. Fig. 5D**).

Lastly, we joined the informative factors of our clinico-pathological parameters with the deep learning model's output to predict survival using a multivariable model. To this end we combined the deep learning model's predicted risk scores and clinico-pathological parameters that showed a p-value below 0.1 in univariate analyses (perineural invasion, invasion beyond subcutaneous tissue, thickness >6mm, ulceration and differentiation grade >1) in a multivariable Cox regression model. This combined model showed that the image data carries more information than the clinico-pathological variables (**global p<0.0001**, **Fig. 5C**). In fact, high-risk classification based on the deep learning model carries a hazard ratio of 28.3 (**multivariable p<0.0001**). In contrast, only perineural invasion remains significant with a hazard ratio of 5.8 (multivariable p=0.02), while the other variables are non-significant (**Fig. 5C**).

Considering that only a fraction of patients is positive for perineural invasion and that clinico-pathological information is frequently incomplete, these analyses highlight the potential of our image-based model to reliably identify patients at high risk of disease progression for intensified clinical follow-up.

**Discussion**

Deep learning has enabled automation of the analysis of large histopathology images. These digital pathology methods not only provide fast and detailed insights into the cellular composition of massive WSIs,[19,20] but also allow to identify patterns and anomalies that may be imperceptible to the human eye.[21] Here we present an approach that combines both: a model that detects complex, imperceptible morphological features of a tumor sample that are predictive of patient outcome with an explainability procedure to disentangle what these features are. While patient outcomes might be influenced by multifactorial clinical variables and span variable development trajectories, we demonstrate that, in case of cSCC, prediction of patient progression is possible based on histological images of their tumor samples alone. Via a comprehensive and quantitative analysis of predictive regions of the tumor samples we point to consistent and repetitive patterns in tumor and tumor microenvironment morphology and organization that characterize progression and non-progression patient groups. Our model offers unmatched accuracy compared to the prediction based on clinico-pathological features that were the gold standard up till now.

Our analysis combined data from three academic clinical centers: Cologne, Munich, and Bonn. The model trained on a single cohort resulted in an uneven accuracy on the remaining two cohorts, ranging from random predictions to 0.84 AUROC. While digital pathology models require large and multi-center data for better generalization, clinical data sharing carries important administrative hurdles and data protection risks. Here we demonstrate that these difficulties can be overcome by employing an FL training scheme resulting in a model with high accuracy across all cohorts while circumventing cross-center data sharing. Our model development strategy allows for easy incorporation of additional clinical centers in the future which could potentially improve the prediction accuracy further.

Deep learning models have achieved human expert-level accuracy in standard diagnostic tasks such as tumor metastases detection and cancer subtyping.[22–24] These tasks involve detecting patterns that, while sometimes local, subtle, and difficult to notice, are known and described in pathology textbooks. In contrast, prediction of patient progression based on WSIs is a more challenging task as there are no known visual biomarkers that reliably indicate disease

advancement. Numerous studies address prediction of cancer progression based on HE-stained samples of tumors across diverse tissue types,[25–30] however rarely reaching accuracy > 0.80 AUROC. Notably, combining image with clinical data has improved prediction accuracy in some studies still barely exceeding 0.80 AUROC.[31–33] In cSCC research, the work of Coudray et al. addresses the prediction of disease outcome from WSIs using a bag of visual words classifier, achieving AUROC=0.689 .[34] These examples demonstrate that prediction of patient progression is indeed difficult, and that the accuracy of our model is among the best achieved so far.

Strikingly, progression risk of a patient could be predicted based on histology images alone, exceeding by far the accuracy achieved by a model trained on clinico-pathological features. Unlike clinical parameters,[7,8] or gene expression measurements,[9,10] which in different clinical centers might follow different standards, be done selectively for some patients only, and come with a high cost, histology is routinely performed in cSCC diagnosis. The fact that tissue slides are available for every patient and that prediction is fast and free of additional costs, considerably increases the facility and potential of our model for clinical use. Moreover, by obviating the need for data sharing, FL greatly facilitates further model training and refinement and its extension to additional centers.

Unlike prediction based on clinical parameters, which are numeric and unambiguous, prediction based on image data is not easy to interpret. Commonly, multiple instance learning models are interpreted using qualitative inspection of image regions with high attention scores.[22–24] Here we adopt a fully quantitative and systematic approach to model interpretation in which we filter predictive patches of each patient group and statistically compare over 524 cell-based features between the two groups. Our features are based on a segmentation model specifically designed for this tumor type and capture a broad range of aspects of sample cell composition, spatial organization of the tissue, as well as nuclei morphology.[17] We point to several noticeable differences in tumor morphology between progressing and non-progressing patients.

Interestingly, the most predictive patches of disease progression were located outside of the tumor region. In contrast, in patients without disease progression, the predictive patches were

inside the tumor according to our IGs-based analysis. On the level of cellular morphology and tissue architecture, tumors from patients with disease progression exhibited a higher degree of heterogeneity. Parameters quantifying nuclear morphology showed higher variability and in these patients, cells in the tumor tissues showed a less uniform distribution. Different areas in and around the cSCC tumor, as well as features of cellular morphology may play distinct roles in the propensity for local recurrence and/or metastatic spread. Future studies in additional cohorts, ideally together with genomic and transcriptomic experiments will be instrumental to further validate our model and infer cause-and-effect relationships between morphological findings and risk of disease progression.

In summary, our study presents an explainable, federated deep learning model that reliably stratifies cSCC patients at high risk of disease progression and identifies their characteristic morphological features. The accuracy, interpretability, and federated implementation of our model hold great promise to better understand the disease and to advance the management of cSCC patients in the future.

## Methods

### Patient cohorts

For the initial training cohort, all patients with a primary cSCC diagnosed and treated by excision at the Department of Dermatology at the University Hospital Cologne (Cologne cohort) between January 2009 to May 2019 were collected. For these patients we used clinico-pathological parameters based on medical records and pathology reports and performed active follow-up regarding disease progression status. In the cohort, 96 patients experienced disease progression (metastasis and/or local recurrence), out of which histological specimens from the primary tumor for 54 patients were available. For the deep learning classification, all available progressors were used together with a random sample of primary tumors from patients without disease progression. Local recurrence or lymph-node/distant metastasis within 2 years after initial diagnosis was considered a progression event. Hematoxylin-Eosin (HE) stained slides obtained during routine work-up of surgical samples were available for 162 patients (progress n=54, non-progress n=108). From the University Hospital Bonn (Bonn cohort) patients diagnosed and treated for cSCC between March 2012 and September 2021 were included. Tumors were excised at the Department of Dermatology or the Department of Oral and Maxillo-facial Surgery and worked up histologically following standard procedures. We identified 23 primary cSCC cases with eventual disease progression (recurrence/metastasis) and randomly selected a group of primary cSCCs without disease progression. Of those, HE slides were available for 39 patients (progress n=23, non-progress n=16). For the cohort from the Department of Dermatology, Technical University Munich (TU Munich, Munich cohort) we identified patients with a primary cSCC and disease progression and assembled a random cohort of primary cSCCs without disease progression. Of those, HE slides were available for 51 patients (progress n=21, non-progress n=30). Patient inclusion and analysis was approved by the institutional review boards (Ethic vote numbers 187/16, 21-1500, 20-1082 and 22-1330-retro).

**Analysis and classification of whole slide images**

**Datasets:** Whole-slide images (WSIs) were acquired from HE slides using a NanoZoomer Slide Scanner (Hamamatsu) at 40x resolution. In total, we collected 219 WSIs of 162 patients from the University Hospital Cologne, 291 WSIs of 39 patients from the University Hospital Bonn, and 129 WSIs of 51 patients from TU Munich. We filtered out slides without any tumor tissue according to the Segmenter model described by Sancéré et al. The final dataset used for training of the federated deep learning model comprises 214 slides from 157 patients from the University Hospital Cologne, 133 slides from 35 patients from the University Hospital Bonn and 113 slides from 51 patients from TU Munich. From this dataset, 228 slides are from patients showing cSCC progression, and 232 slides are from patients showing no cSCC progression. Data splitting is done in a stratified fashion on patient level, making 65-15-20 splits for training, validation, and testing, respectively.

**Pre-processing:** Each WSI is tiled into patches of 256x256 pixels at x20 magnification. Patches without tissue are discarded, and the remaining patches are processed with an ImageNet pre-trained EfficientNet-v2-L,[35] to compute its feature vector representations.

**Classification:** Each WSI is treated as the sequence of feature vectors corresponding to its non-empty image patches. We use the multiple instance learning classification model described by Pisula and Bozek.[36] Following an approach similar to Lu et al.,[37] a transformer model initialized with language-modeling pre-training weights is used for classification. We use a RoBERTa transformer encoder,[38] and perform parameter-efficient fine-tuning by only training its normalization layers.[37,39] To reduce compute and memory footprint, we apply multi-head attention pooling at the input to shorten the length of the patch sequence. The embedding vectors from the last layer of the transformer encoder are averaged and fed to a linear layer for the final classification.

Each WSI is classified independently during model training. During inference, in cases where there are multiple slides per patient, we evaluate the model on each one and take the prediction corresponding to the slide with the biggest activation in the positive class output neuron.

**Model training:** We train our model with a Federated Averaging strategy for 50 rounds.[12] Adam is used as the optimizer algorithm, with a learning rate of 1.e-4, weight decay of 5.e-5,

and batch size of 4. Model selection is done based on weighted validation AUROC of the three cohorts.

**Classification explanation and analysis**

Beyond mere disease progression prediction with a deep network classifier, we investigate the biological features that drive our classifier's decision. Our process is threefold: we detect relevant image regions responsible for the model's decision; we compute handcrafted features of the cellular composition of the image regions; and we perform the data analysis itself. This approach is described in detail below.

**Input attributions**

We use Integrated Gradients (IGs) to identify regions of a WSI that play a role in the classifier's progression prediction.[16] IGs is a deep learning explainability algorithm that attributes the prediction of a deep network to its input features. We apply IGs to our cSCC progression prediction model, to assign a positive score to image patches that contribute to the prediction of the correct class, and a negative score to patches that contribute to the prediction of the opposite outcome. By arranging the IGs attribution scores of the patches in their corresponding spatial locations in the slides, it is possible to visualize these values as heatmaps, as shown in **Fig. 3**.

**Patch description and feature engineering**

We use the HoverNet nuclei segmentation model described by Sancéré et al. on the WSI image patches to identify their cell composition.[17,19] The model detects and classifies cell nuclei into granulocytes, lymphocytes, plasma cells, stroma cells, tumor cells, and non-neoplastic epithelial cells. Once the cells in a patch have been identified, we compute a total of 524 features that summarize the patch into a single feature vector. These features include:

- Cell type populations and ratios.
- Descriptive statistics (mean, median, variance, skewness, kurtosis, minimum, maximum) of nuclei morphology, such as the mean tumor cells nuclei eccentricity, or the variance in plasma cells nuclei area. These features were computed with the `skimage.measure` Python package.[40]
- Descriptive statistics of distances between cell nuclei, such as the median distance between stroma cells and tumor cells.

- Average Nearest Neighbor Ratio (ANNR) and Join Count analysis for each cell type.

The features from the last item are used to quantify the spatial arrangement of cells within a patch, and they capture two different aspects of it.

ANNR is used to quantify the observed pattern of distances between cell nuclei in a patch:

$$ANNR = \underline{D_O} \,/\, \underline{D_E},$$

where $\underline{D_O}$ is the observed mean distance between each cell and its closest neighbor, and $\underline{D_E}$ is the expected mean distance between each cell and its closest neighbor if the cells were placed randomly:

$$\underline{D_E} = \frac{0.5}{\sqrt{n/A}},$$

where $n$ is the number of cells in a patch, and $A$ is the patch area. An ANNR<1 indicates clustering (meaning, cells in the patch are closer than a random pattern of cells), and an ANNR>1 indicates a dispersed or regular pattern of cell nuclei. We compute the ANNR for each cell type in a patch.

Join Count analysis gives a measure of spatial autocorrelation: it describes how the values of a variable at neighboring spatial locations are similar to each other. In our case, the variable of interest is the cell type, where a positive spatial autocorrelation would mean that neighboring cells belong to the same type, and a negative spatial autocorrelation would mean that neighboring cells belong to different classes. Spatial autocorrelation is complementary to ANNR, it quantifies neighboring cell nuclei types disregarding how close or distanced they are. Our Join Count analysis is computed for each cell type individually, in the following way:

- A patch is partitioned into a Voronoi tessellation, using the nuclei centroids as seeds for the regions.
- The regions are binary-labeled. Given a cell type, a positive label is assigned to all the cell nuclei belonging to that class, and a negative label is assigned to the remaining regions.
- The different types of joins were then counted. Two neighboring cells make a black-black (BB) join if they both are from the positive label (i.e. the cell type being currently analyzed); a black-white (BW) join is formed between two cells of opposite labels; and a white-white (WW) join happens when two cells of the negative label neighbor each other.

This procedure is done for each cell type independently, assigning the positive label (black) to the analyzed cell type and the negative label (white) to all the other cell types. Our measure of spatial autocorrelation is given by:

$$Spatial\ Autocorrelation\ = (J_{BB}\ -\ J_{BW})\ /\ J_T\ ,$$

where $J_{BB}$ , $J_{BW}$ , and $J_T$ are the number of BB joins, the number of BW joins, and the total number of joins, respectively. This equation is positive when the majority of joins in a patch are BB joins, indicating a positive spatial autocorrelation, and is negative when the majority of joins are BW joins, indicating negative spatial autocorrelation.

**Data analysis**

We apply IGs to all the patients in the test set, and describe their corresponding image patches as previously explained. We use in this analysis the patches coming from tumor regions detected by the Segmenter model described by Sancéré et al.,[17,41] plus a surrounding tissue stripe of approximately 800µm of width next to the tumor border. From the totality of patches, we form two groups: A "positive group" of image patches coming from progression patients, which were detected to be explainable of this condition with IGs; and a "negative group" of patches coming from non-progression patients, which were detected to be explainable of this condition with IGs.

To enhance the predictive signal and avoid over-representing patients with bigger tumors, we take a slide's top 10% IGs-scored patches, and limit this quantity to 200 image patches per slide. We compare values of each feature individually between the two groups of patches. We guide our analysis by focusing on features whose values differ between the two groups with an Effect Size bigger than random. We use the Common Language Effect Size (CLES),[42] or probability of superiority, as it has no assumptions about the data distribution, and is straightforward to understand:

$$CLES\ =\ P(X > Y)\ ,$$

is the probability that a value sampled from group X is bigger than a value sampled from group Y. In our case, the two groups are the positive and the negative groups previously described, and we compute the CLES for each feature with brute force, by exhaustively comparing each value of one group with all the values of the same feature in the other group.

In addition to comparing the feature distributions in both groups, we tested whether the individual patches' feature vectors were sufficient to predict the progression status of their

respective patients using an XGBoost classifier.[18] The patches under analysis were split into 80-20 train and test sets, and model selection was done with 3-fold cross-validation on the train set.

**Statistical analysis of clinico-pathological variables**

Associations of clinico-pathological variables with disease progression and survival were done for all patients with available data. Association with disease progression risk was calculated using logistic regression and reported as odds ratios. Association with survival was done using the Kaplan-Meier method with log-rank test as well as Cox proportional hazard models and reported as hazard ratios with 95% confidence intervals. For multivariable analyses, variables with $p<0.1$ in univariate analysis were combined. Analyses were done in R statistical environment (v4.3.0).

**Acknowledgements**

A.F. was partly funded by the Deutsche Krebshilfe through a Mildred Scheel Foundation Grant (grant number 70113307). C.L. was partly funded through the collaborative research center grant on small cell lung cancer (CRC1399, project ID 413326622) by the German Research Foundation (DFG). Both K.B. and J.I.P. were hosted by the Center for Molecular Medicine Cologne throughout this research. K.B. and J.I.P. were supported by the BMBF program Junior Group Consortia in Systems Medicine (01ZX1917B) and BMBF program for Female Junior Researchers in Artificial Intelligence (01IS20054).

**Conflict of Interests**

D.N. received financial support (speaker's honoraria, advisory boards, travel expense reimbursements or grants) from Abbvie, Almirall, AstraZeneca, Biogen, Boehringer Ingelheim, Bristol-Myers-Squib, GlaxoSmithKline, Incyte, Janssen-Cilag, Kyowa Kirin, LEO Pharma, Lilly, L'Oreal/Cerave, MSD, Novartis, Pfizer, Regeneron and UCB Pharma. J.B. received research funding from Bayer outside the presented work. K.D. received financial support (speaker's honoraria, advisory boards, travel expense reimbursements or grants) from Abbvie, Bristol-Myers-Squib, Novartis, and Pierre-Fabre.

**References**


1 Winge MCG, Kellman LN, Guo K, *et al.* Advances in cutaneous squamous cell carcinoma. *Nat Rev Cancer* 2023; **23**: 430–49.

2 Keim U, Katalinic A, Holleczek B, Wakkee M, Garbe C, Leiter U. Incidence, mortality and trends of cutaneous squamous cell carcinoma in Germany, the Netherlands, and Scotland. *Eur J Cancer Oxf Engl 1990* 2023; **183**: 60–8.

3 Brantsch KD, Meisner C, Schönfisch B, *et al.* Analysis of risk factors determining prognosis of cutaneous squamous-cell carcinoma: a prospective study. *Lancet Oncol* 2008; **9**: 713–20.

4 Schmults CD, Karia PS, Carter JB, Han J, Qureshi AA. Factors predictive of recurrence and death from cutaneous squamous cell carcinoma: a 10-year, single-institution cohort study. *JAMA Dermatol* 2013; **149**: 541–7.

5 Thompson AK, Kelley BF, Prokop LJ, Murad MH, Baum CL. Risk Factors for Cutaneous Squamous Cell Carcinoma Recurrence, Metastasis, and Disease-Specific Death: A Systematic Review and Meta-analysis. *JAMA Dermatol* 2016; **152**: 419–28.

6 Eigentler TK, Dietz K, Leiter U, Häfner H-M, Breuninger H. What causes the death of patients with cutaneous squamous cell carcinoma? A prospective analysis in 1400 patients. *Eur J Cancer* 2022; **172**: 182–90.

7 Ruiz ES, Karia PS, Besaw R, Schmults CD. Performance of the American Joint Committee on Cancer Staging Manual, 8th Edition vs the Brigham and Women's Hospital Tumor Classification System for Cutaneous Squamous Cell Carcinoma. *JAMA Dermatol* 2019; **155**: 819–25.

8 Schmults CD, Blitzblau R, Aasi SZ, *et al.* NCCN Guidelines® Insights: Squamous Cell Skin Cancer, Version 1.2022. *J Natl Compr Cancer Netw JNCCN* 2021; **19**: 1382–94.

9 Wysong A, Newman JG, Covington KR, *et al.* Validation of a 40-gene expression profile test to predict metastatic risk in localized high-risk cutaneous squamous cell carcinoma. *J Am Acad Dermatol* 2021; **84**: 361–9.

10 Wang J, Harwood CA, Bailey E, *et al.* Transcriptomic analysis of cutaneous squamous cell carcinoma reveals a multigene prognostic signature associated with metastasis. *J Am Acad Dermatol* 2023; **89**: 1159–66.

11 Haenssle HA, Fink C, Schneiderbauer R, *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol Off J Eur Soc Med Oncol* 2018; **29**: 1836–42.

12 McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. 2023; published online Jan 26. http://arxiv.org/abs/1602.05629 (accessed June 21, 2024).

13 Ogier Du Terrail J, Leopold A, Joly C, *et al.* Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat Med* 2023; **29**: 135–46.

14 Maron O, Lozano-Pérez T. A Framework for Multiple-Instance Learning. In: Advances in Neural Information Processing Systems. MIT Press, 1997. https://proceedings.neurips.cc/paper_files/paper/1997/hash/82965d4ed8150294d4330ace00821d77-Abstract.html (accessed June 24, 2024).

15 Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need. 2023; published online Aug 1. DOI:10.48550/arXiv.1706.03762.
16 Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. 2017; published online June 12. http://arxiv.org/abs/1703.01365 (accessed May 22, 2023).
17 Sancéré L. Histo-Miner: Tissue Features Extraction With Deep Learning from H&E Images of Squameous Cell Carcinoma Skin Cancer. Manuscript in preparation. https://github.com/bozeklab/histo-miner (accessed July 15, 2024).
18 Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001; **29**: 1189–232.
19 Graham S, Vu QD, Raza SEA, *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal* 2019; **58**: 101563.
20 Hörst F, Rempe M, Heine L, *et al.* CellViT: Vision Transformers for Precise Cell Segmentation and Classification. 2023. DOI:10.48550/ARXIV.2306.15350.
21 Kather JN, Pearson AT, Halama N, *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25**: 1054–6.
22 Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5**: 555–70.
23 Shao Z, Bian H, Chen Y, *et al.* TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., 2021: 2136–47.
24 Chen RJ, Chen C, Li Y, *et al.* Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. 2022; published online June 6. DOI:10.48550/arXiv.2206.02647.
25 Dietrich E, Fuhlert P, Ernst A, *et al.* Towards Explainable End-to-End Prostate Cancer Relapse Prediction from H&E Images Combining Self-Attention Multiple Instance Learning with a Recurrent Neural Network. In: Proceedings of Machine Learning for Health. PMLR, 2021: 38–53.
26 Akram F, Wolf JL, Trandafir TE, Dingemans A-MC, Stubbs AP, von der Thüsen JH. Artificial intelligence-based recurrence prediction outperforms classical histopathological methods in pulmonary adenocarcinoma biopsies. *Lung Cancer* 2023; **186**: 107413.
27 Wu Z, Wang L, Li C, *et al.* DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images. *Front Genet* 2020; **11**. DOI:10.3389/fgene.2020.00768.
28 Xiao H, Weng Z, Sun K, *et al.* Predicting 5-year recurrence risk in colorectal cancer: development and validation of a histology-based deep learning approach. *Br J Cancer* 2024; **130**: 951–60.
29 Foersch S, Glasner C, Woerl A-C, *et al.* Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med* 2023; **29**: 430–9.
30 Shi Y, Olsson LT, Hoadley KA, *et al.* Predicting early breast cancer recurrence from histopathological images in the Carolina Breast Cancer Study. *Npj Breast Cancer* 2023; **9**: 1–7.
31 Howard FM, Dolezal J, Kochanny S, *et al.* Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *Npj Breast Cancer* 2023; **9**: 1–6.
32 Yang J, Ju J, Guo L, *et al.* Prediction of HER2-positive breast cancer recurrence and

metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput Struct Biotechnol J* 2022; **20**: 333–42.
33 Lucas M, Jansen I, van Leeuwen TG, Oddens JR, de Bruin DM, Marquering HA. Deep Learning–based Recurrence Prediction in Patients with Non–muscle-invasive Bladder Cancer. *Eur Urol Focus* 2022; **8**: 165–72.
34 Coudray N, Juarez MC, Criscito MC, *et al.* Self-supervised artificial intelligence predicts recurrence, metastasis and disease specific death from primary cutaneous squamous cell carcinoma at diagnosis. *Res Sq* 2023; : rs.3.rs-3607399.
35 Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv* 2019; published online May 24. https://www.semanticscholar.org/paper/EfficientNet%3A-Rethinking-Model-Scaling-for-Neural-Tan-Le/4f2eda8077dc7a69bb2b4e0a1a086cf054adb3f9 (accessed June 21, 2024).
36 Pisula JI, Bozek K. Language models are good pathologists: using attention-based sequence reduction and text-pretrained transformers for efficient WSI classification. 2023; published online Sept 30. DOI:10.48550/arXiv.2211.07384.
37 Lu K, Grover A, Abbeel P, Mordatch I. Pretrained Transformers as Universal Computation Engines. 2021; published online June 30. http://arxiv.org/abs/2103.05247 (accessed June 6, 2024).
38 Liu Y, Ott M, Goyal N, *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019; published online July 26. DOI:10.48550/arXiv.1907.11692.
39 Lialin V, Deshpande V, Rumshisky A. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. 2023; published online March 27. http://arxiv.org/abs/2303.15647 (accessed June 6, 2024).
40 Walt S van der, Schönberger JL, Nunez-Iglesias J, *et al.* scikit-image: image processing in Python. *PeerJ* 2014; **2**: e453.
41 Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for Semantic Segmentation. 2021; published online Sept 2. DOI:10.48550/arXiv.2105.05633.
42 McGraw KO, Wong SP. A common language effect size statistic. *Psychol Bull* 1992; **111**: 361–5.

Chapter 5

# Efficient WSI classification with sequence reduction and transformers pre-trained on text

In the previous chapter, a neural network that was pre-trained to model language in a large text dataset was used to predict progression risk from pathology slide images of CSCC patients. While this approach may sound flawed, given that the pre-training data is out of domain (and out of modality!), there exists an astounding hypothesis: text-pre-trained transformers can transfer well to other tasks outside the world of natural language processing.

This chapter explores the classification of histopathology WSIs with pre-trained language models. A natural challenge arises given that WSIs constitute a prohibitively large input to deep attention models whose time and space complexity is $O(n^2)$. An attention-based layer to pool visual information into a compact sequence is then developed, allowing the processing of arbitrarily large pathology slides with any deep transformer model. The memory footprint of the algorithm is reduced even more by only training the normalization layers of the model.

scientific reports

Check for updates

OPEN

# Efficient WSI classification with sequence reduction and transformers pretrained on text

Juan I. Pisula[1,2]✉ & Katarzyna Bozek[1,2,3]

**From computer vision to protein fold prediction, Language Models (LMs) have proven successful in transferring their representation of sequential data to a broad spectrum of tasks beyond the domain of natural language processing. Whole Slide Image (WSI) analysis in digital pathology naturally fits to transformer-based architectures. In a pre-processing step analogous to text tokenization, large microscopy images are tessellated into smaller image patches. However, due to the massive size of WSIs comprising thousands of such patches, the problem of WSI classification has not been addressed via deep transformer architectures, let alone via available text-pre-trained deep transformer language models. We introduce SeqShort, a multi-head attention-based sequence shortening layer that summarizes a large WSI into a fixed- and short-sized sequence of feature vectors by removing redundant visual information. Our sequence shortening mechanism not only reduces the computational costs of self-attention on large inputs, it also allows to include standard positional encodings to the previously unordered bag of patches that compose a WSI. We use SeqShort to effectively classify WSIs in different digital pathology tasks using a deep, text pre-trained transformer model while fine-tuning less than 0.1% of its parameters, demonstrating that their knowledge about natural language transfers well to this domain.**

Tranformers[1] have brought several breakthroughs to the disciplines of natural language processing (NLP) and computer vision (CV). Their capacity to link information across sequences of vector embeddings, representing either visual features or vectorized words, allowed to capture the structure and meaning necessary for machine translation[2–5], question-answering[6–9], image classification[10–12] and segmentation[11,13], and even multi-modal tasks such as text-to-image generation[14,15].

Concurrently in the field of digital pathology, the popularization of Multiple Instance Learning (MIL)[16,17] approaches for Whole Slide Image (WSI) analysis allowed for the fast adoption of transformer models in this domain. By considering each WSI as a set of feature vectors of smaller tissue patches, this type of data is a natural input to transformer architectures. However, although transformer-based, these methods are typically modified and adapted to the idiosyncrasies of MIL and histopathology. Given gigapixel image size, out-of-the-box Vision Transformers (ViTs)[10] are excessively memory-demanding. Diverse shapes of WSIs and removal of patches consisting of background, artifacts, such as pen marker lines, require tailored implementation of local or windowed attention[18,19]. Novel positional encoding methods have been proposed to replace fixed and learnable positional embeddings commonly found in NLP transformers or ViTs[20–25]. To overcome the challenges of WSI processing, we base our work on the two observations below.

- *The redundancy of information present in full-sequence self-attention operations* can be exploited to reduce the computational cost of large inputs in deep transformer models. Wang *et al.*[26] base their Linformer model on the observation that an attention matrix can be approximated with a matrix of lower rank. The works of Liu *et al.*[11] and Dai *et al.*[27] propose to construct hierarchical representations instead of maintaining full-length, token-level resolution. The observations made by Clark *et al.*[28] about the importance of the [SEP] token and neighboring tokens have inspired several methods of local and sparse attention[29–32]. Comprising thousands of image patches, a WSI representation in a MIL approach is a prohibitively long sequence of vector embeddings. *We hypothesise that such findings in the transformer literature are valid to histopathology data as well*, and techniques for attention matrix reduction are necessary to allow for processing of massive in size WSIs with the use of transfomers.

[1]Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [2]Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [3]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany. ✉email: juan.pisula@uk-koeln.de

- *Text pre-trained transformers have been proven successful in non-language related tasks.* Recent works have shown that language models pre-trained on large unstructured text corpora not only perform strongly in various downstream NLP tasks, but in several tasks outside of this domain, ranging from solving math problems[33], to lossless image and audio compression[34]. We refer to[35] for an extensive enumeration of such works. In the context of CV, Ilarco *et al.*[36] showed that text representations of frozen language models are predictive of visual representations of their corresponding object. More recently, Lu *et al.*[37] demonstrated that pre-trained language models show high performance in image classification, numerical computation, and protein fold prediction when less than 0.1% of their parameters are fine-tuned. *Language-based pre-training can therefore be leveraged to perform different, out-of-domain tasks,* which however has never been demonstrated in WSI classification.

In this work we use deep transformer architectures to classify WSI data. To allow for processing of thousands of image patches from a single slide, we propose *SeqShort*, a multi-head attention (MHA) input layer that reduces long input sequence to a fixed-size short sequence that can be processed by any transformer model. Furthermore, we show that classification performance is increased when the transformer classifier is pre-trained with a language modeling task compared to training it from scratch, and that only fine-tuning less than 0.1% of its weights is necessary. This way, we construct a deep, yet computationally inexpensive model that requires a reduced set of trainable parameters, and performs well in digital pathology tasks.

## Results

We compress the visual information of WSIs with our sequence reduction technique and use transformer models trained from scratch or pre-trained on text data to solve several WSI classification tasks. We train multiple transformer architectures and find that text pre-training improves classification performance in deep transformer models. In our approach the input is in a form of an ordered sequence, instead of an unordered collection of image patches as commonly done in other MIL algorithms. We further show that positional information that we add to the ordered sequences is taken into account by the transformer classifier and improves its prediction accuracy.

We then examine how our SeqShort layer works to better understand how visual information in the WSIs is aggregated. We find that only a small subset of image patches per WSI is relevant to produce their compressed sequence representations, corroborating our hypothesis about information redundancy in WSIs. Although these representations act as potentially lossy summaries of the WSIs, an extension of the attention rollout algorithm[38] can trace the output of the transformer classifiers back to each individual image patch, providing an interpretability mechanism for the classification outcome.

### WSI classification

We measure the performance of our method on three different classification tasks: Lymph Node Metastases (LNM) classification (Normal vs Metastases); Invasive Breast Carcinoma (IBC) subtype classification (Invasive Ductal Carcinoma vs Invasive Lobular Carcinoma); and Renal Cell Carcinoma (RCC) subtype classification (Papillary Cell Carcinoma vs Chromophobe Cell Carcinoma vs Clear Cell Carcinoma). For the LNM classification task we use the dataset provided by the CAMELYON16 grand challenge (https://camelyon16.grand-challenge.org/), keeping 10% of the training samples as a validation set, and evaluating on the grand challenge test set. For the cancer subtyping tasks, we use WSIs collected from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga), and follow the same stratified 10-fold cross-validation as[40,44].

We use 256×256 image patches cropped from the WSIs both at ×10 and ×20 magnification. As a data scarcity ablation, we train the models using the complete datasets or just 25% of the samples. Area under ROC curve (AUROC) is used as classification performance metric. We compare our method against several state-of-art weakly supervised architectures. All networks are compared using a single magnification at a time, and are agnostic of how their input features vectors were produced. We use an EfficientNetV2-L[43] pre-trained on ImageNet[45] as patch-level feature extraction network in this work. As our best performing model we use a frozen RoBERTa-base[8] model as MIL classifier, and only fine-tune its normalization layers. Results of this experiment are shown in Table 1, and additional results of LNM and IBC classification using a CTransPath[46] feature extractor are shown in the Supplementary Table 1.

Although 99.9% of the parameters in our model were trained solely on text data, it surpasses WSI-specific methods in most LNM and IBC experiments, while demonstrating competitive AUROC in RCC classification. The experiments utilizing the CTransPath feature extractor generally enhance the performance of all models, with a slight performance decline for our model only in the LNM task when using 25% of the data.

### Pre-training on text improves WSI classification

We explore the use of popular NLP transformer architectures that can be trained in a single GPU for sequence classification in the WSI classification task. Such architectures have not been applied before in weakly-supervised histopathology tasks given the computational cost of handling thousands of instances in a single WSI. The sequence shortening method that we introduce in this study allows us to overcome the computational cost problem.

Inspired by the success of pre-trained language models in different tasks outside the NLP, we propose the use of a frozen, language-modeling pre-trained transformer as MIL classifier. This is motivated by the hypothesis that the multi-head self-attention (MHSA) layers of a transformer language model learn to capture the interdependencies among the elements of sequences, independent of the original data modality or domain. We follow[37] and only fine-tune the normalization layers of the model, reducing the amount of trainable parameters in our transformer encoder from 85M to 36K (only 0.04% of the total amount).

| Method | x10 magnification | | x20 magnification | |
|---|---|---|---|---|
| | 25% train set | 100% train set | 25% train set | 100% train set |
| Lymph Node Metastases classification | | | | |
| ABMIL[39] | 0.501 | 0.664 | *0.516* | 0.616 |
| CLAM[40] | 0.511 | 0.692 | *0.516* | 0.673 |
| DS-MIL[41] | 0.468 | 0.695 | 0.441 | 0.640 |
| TransMIL[20] | *0.529* | 0.629 | 0.470 | 0.723 |
| Wagner et al.[42] | 0.465 | **0.778** | 0.501 | *0.778* |
| Ours | **0.627** | *0.772* | **0.642** | **0.865** |
| Invasive Breast Carcinoma subtype classification | | | | |
| ABMIL[39] | 0.542 ± 0.107 | 0.571 ± 0.088 | 0.551 ± 0.103 | 0.554 ± 0.107 |
| CLAM[40] | 0.811 ± 0.055 | 0.850 ± 0.039 | 0.697 ± 0.056 | 0.791 ± 0.082 |
| DS-MIL[41] | 0.779 ± 0.075 | 0.892 ± 0.045 | 0.711 ± 0.084 | 0.819 ± 0.082 |
| TransMIL[20] | *0.864 ± 0.063* | *0.896 ± 0.048* | **0.782 ± 0.094** | *0.856 ± 0.064* |
| Wagner et al.[42] | 0.687 ± 0.202 | 0.854 ± 0.069 | 0.739 ± 0.099 | 0.824 ± 0.077 |
| Ours | **0.874 ± 0.052** | **0.901 ± 0.049** | *0.765 ± 0.099* | **0.863 ± 0.047** |
| Renal Cell Carcinoma subtype classification | | | | |
| ABMIL[39] | 0.724 ± 0.077 | 0.795 ± 0.040 | 0.697 ± 0.077 | 0.758 ± 0.044 |
| CLAM[40] | **0.965 ± 0.013** | 0.969 ± 0.025 | *0.961 ± 0.013* | 0.974 ± 0.010 |
| DS-MIL[41] | 0.941 ± 0.047 | 0.971 ± 0.001 | 0.926 ± 0.025 | 0.963 ± 0.001 |
| TransMIL[20] | *0.962 ± 0.015* | **0.980 ± 0.001** | **0.971 ± 0.010** | *0.980 ± 0.001* |
| Wagner et al.[42] | 0.960 ± 0.020 | *0.979 ± 0.009* | 0.971 ± 0.011 | **0.984 ± 0.007** |
| Ours | 0.942 ± 0.019 | 0.974 ± 0.011 | 0.952 ± 0.017 | 0.977 ± 0.013 |

**Table 1**. Performance of different MIL algorithms in the different slide-level classification tasks using EfficientNet features[43]. Best and the second best classification results are in bold and italics, respectively.

| Language Model | AUROC |
|---|---|
| Baseline | 0.784 ± 0.082 |
| XLNet-base[47] | 0.819 ± 0.090 |
| GPT2-small[48] | 0.827 ± 0.079 |
| BERT-base[6] | 0.849 ± 0.058 |
| ALBERT-base[7] | 0.747 ± 0.118 |
| Llama3-8B[49] | 0.810 ± 0.070 |
| RoBERTa-base[8] | **0.863 ± 0.047** |

**Table 2**. Performance of different Language Models in IBC subtype classification, at $\times 20$ magnification.

An important question is if text pre-training does play a role in classification performance. We compare the performance of a baseline transformer encoder trained from scratch with different frozen text-pre-trained transformers. Given our GPU memory constrains, the SeqShort layer was required in order to train these models. All the tested models have 12 layers of 12 attention heads and 768 hidden units, resulting in comparable transformer size across all models, except for Llama3-8B[49]. Llama3-8B is a larger, 8 billion parameter model comprising 32 layers of 32 attention heads and 4096 hidden units. The model was fine-tuned with 8-bit model weights to fit it in our hardware. The baseline model, BERT-base, and RoBERTa-base have identical architecture and only differ in text-pre-training dataset and language modeling task.

This experiment was done at $\times 20$ magnification, using the IBC dataset. Except for ALBERT-base[7], every model outperforms the baseline (Table 2) indicating that pre-training on large corpus of text data does influence model performance in other domains including digital pathology.

## The role of positional information

In our approach, we consider a WSI to be an unordered bag of image patches, and SeqShort provides positional information for free by reducing it to an ordered, fixed-length sequence of feature vectors. This enables the adding of the fixed-size set of learnable positional embeddings which is common practice in transformer architectures of CV and NLP tasks to the output of SeqShort.

Different positional encodings based on patch location have been proposed to address the problem of varying WSI shapes and sizes[20–25], and their inclusion is compatible with our method. In this section, we repeat the IBC subtyping experiment to investigate the effect of positional information on classification performance. In this

experiment, we enhance our classifier with the patch location positional encoding used in[23] previous to the SeqShort input, in addition to the standard BERT positional embedding used before the transformer classifier.

The results of the experiment are shown in Table 3. In the unordered bag of patches formulation, the ordered output sequence of SeqShort carries positional information that can be exploited by adding positional encoding, increasing AUROC by 0.017 at $\times 10$ magnification and by 0.038 at $\times 20$ magnification. Positional encoding of the patches based on their 2D spatial location also improves performance, and the best results are achieved when both types of positional encoding are employed.

### Insights into sequence summarization

We probe the SeqShort layer to examine how a WSI is summarized. We calculate the Kullback-Leibler (KL) divergence between the attention distributions produced by the different learned query vectors in SeqShort and a uniform attention distribution. Values close to zero indicate that such queries pay overall the same amount of attention to all the input patches, whereas higher KL divergence values suggest that such queries pay more attention to a reduced subset of image patches. We do this measurement with every sample of one of the IBC test sets at $\times 20$ magnification, and average the results.

The KL divergence values are shown in Fig. 2, as well as an example WSI and the attention heatmaps produced by three different learned query vectors. The individual heatmaps demonstrate that indeed some patches receive more attention than the others. However across the three heatmaps, even though the attention distributions are spread over various-sized image areas, the same patches receive high-attention.

We confirm this visual insight by calculating the Spearman's rank correlation coefficients between pairs of different learned query vectors' ranking of patches (within a single WSI). For the WSI in Fig. 2 and the three examined query vectors, the correlation coefficients are above 0.96, and when considering the complete set of 256 query vectors, the mean rank correlation coefficient value is 0.99 (with a minimum value of 0.76). Among all the WSIs in the test set, 99.7% of the total pairs of rankings show a correlation coefficient > 0.7.

### Explanation of classification outcome

Attention heatmaps from the previous experiment illustrate the functioning of the SeqShort layer of our model: they provide insights into how the individual patches of a WSI are weighed to synthesize the intermediate output of our method.

Given how the model aggregates the patch representations throughout its forward pass, we apply attention rollout[38] to generate heatmaps that provide insights into the overall attention the model assigns to each patch in its decision process. We modify the base case of the recursive definition of attention rollout to take into consideration that SeqShort is the first layer of the complete model. Our modified attention rollout is then defined as:

$$\tilde{\boldsymbol{A}}_i = \begin{cases} \boldsymbol{A_i} \cdot \tilde{\boldsymbol{A}}_{\boldsymbol{i-1}} & \text{if } i > 0 \\ \begin{bmatrix} \mathbf{0} \\ \boldsymbol{A_i} \end{bmatrix} & \text{if } i = 0, \end{cases} \tag{1}$$

where $\boldsymbol{A_i}$ is the attention matrix of layer $i$, and $\mathbf{0}$ is the zero vector in row space, to take into account that the [CLS] token was not present in the MHA operation of SeqShort. Example heatmaps are shown in Fig. 3. Hence, while allowing to process large WSIs, the SeqShort mechanism does not limit the interpretability of the predictive model.

## Discussion

In this work we use a text pre-trained transformer model for WSI classification. Such pre-training has been shown to transfer to other modalities, and we corroborate this finding in three digital pathology tasks. To do so, we use a standalone layer for sequence reduction aimed to overcome common challenges in WSI classification with transformer architectures, and to reduce the compute budget required for processing large inputs with deep self-attention-based architectures.

| Magnification | Pos. embedding | | AUROC |
|---|---|---|---|
| | WSI | Seq. | |
| $\times 20$ | No | No | 0.825 ± 0.052 |
| | No | Yes | 0.863 ± 0.047 |
| | Yes | No | 0.865 ± 0.044 |
| | **Yes** | **Yes** | **0.866 ± 0.064** |
| $\times 10$ | No | No | 0.884 ± 0.062 |
| | No | Yes | 0.901 ± 0.049 |
| | Yes | No | 0.916 ± 0.046 |
| | **Yes** | **Yes** | **0.917 ± 0.035** |

**Table 3**. Effect of including positional information on classification performance of IBC subtyping.

Our SeqShort layer was developed with the hypothesis that there is redundant visual information in the full sequence of patches of a WSI, similar to the redundancy in text sequences previously explored in the NLP literature[26–32]. Our results in section 2.4 show that high-attention patches are preserved throughout the different learned queries, indicating their importance for the prediction. Indeed 99.7% of the pairs of patch rankings based on different query vectors show correlation coefficient $> 0.7$. These results suggest that there is redundancy in the full sequence of patches, as certain patches are consistently more important than others, and that classification is possible by aggregating them into a shorter sequence. Moreover, in section 2.3, we show that classification performance is increased between 0.017 and 0.038 when the downstream classifier is augmented with positional embeddings that encodes the sequential order in the output generated by the SeqShort layer. For the sake of simplicity, most of our experiments are done considering a WSI to be an unordered bag of image patches. However, including patch location positional encoding is compatible with our approach, producing a further performance boost.

Text pre-trained transformers have been proven successful in non-language related tasks[33,35–37]. In section 2.2 we show that classification performance can be increased by up to 0.079 AUROC points just by fine-tuning less than 0.1% of the parameters of a deep transformer that was pre-trained on a large text dataset, compared to the same model trained from scratch. The best performing model in our experiments is RoBERTa-base, which outperformed BERT-base in the WSI classification, reflecting these models' performance difference in several NLP tasks. Notably, these models have the same architecture but differ in the pre-training objective and dataset size. Only the ALBERT-base LM was outperformed by the model trained from scratch. In contrast to the rest of the models in this experiment, ALBERT-base contains a single fully trainable layer whose parameters are reused in the subsequent layers, which might explain its lower capacity of transferring to other domains. These results suggest that both the transformer size and text corpus volume play a role in the model performance in a WSI classification task.

Our primary goal is not to design a novel MIL algorithm that surpasses state-of-the-art, but rather to demonstrate that out-of-the-box LMs can transfer their representations of sequential data to the field of digital pathology. Models designed for this discipline are very performant, have a parameter count orders of magnitude smaller than LMs, and inference time considerably faster. We consider it a reasonable decision to employ WSI MIL classifiers instead of models that were designed and trained for NLP. In section 2.1 our LM-based approach outperforms the WSI-dedicated methods in most LNM and IBC experiments, and showing competitive AUROC in RCC classification. The experiments using the CTransPath feature extractor show a general increase of performance for all models, and is only detrimental for our model when using 25% of the data in the LNM task. These results show that LMs are competitive WSI classifiers, outperforming MIL models in some of the tasks, and suggesting that this direction of research in digital pathology is worth exploring further.

The scope of this work is limited to "base" LMs that comprise 80 million parameters, and are possible to fit in a single GPU. We included an experiment with Llama3-8B[49] in table 2. With 8 billion parameters, we could only fine-tune it using 8-bit quantized model weights, making this experiment not directly comparable to the rest of the models in the comparison. A natural extension of our study is to do further experiments with Large Language Models such as OPT-175B, with 175 billion parameters[50], or the rest of the Llama family of models that comprise up to 405 billion parameters[49].

## Methods

### Sequence shortening

Existing methods[11,27,51] for sequence reduction are not suitable for MIL WSI problems. Since there is no spatial information about instances in an unordered bag, concatenating neighboring feature vectors or taking their strided average is meaningless, as the order of the patches in a bag is arbitrary. Methods that employ a linear projection for dimensionality reduction after instance concatenation or sequence reshaping are not applicable to WSIs either, as they require a fixed and known input shape.

Here we propose using MHA for sequence shortening. Similar ideas have been explored in text-vision multi-modal understanding tasks[52–54], and is reminiscent of how object queries are used in transformer object detection[55], with the advantage of not requiring object-level annotations.

Given $\boldsymbol{X} \in \mathbb{R}^{M \times d}$ the sequence of $M$ $d$-dimensional feature vectors of non-overlapping WSI tiles, we introduce our *SeqShort* input layer that generates a new sequence $\boldsymbol{X_S} \in \mathbb{R}^{S \times h}$ with an MHA layer:

$$\begin{aligned} \boldsymbol{X_S} = & \text{ MHA } (Q = \boldsymbol{Q_l}, K = \boldsymbol{X}, V = \boldsymbol{X}) + \boldsymbol{Q_l} \\ = & \text{ Concat } (head_1, ..., head_k)\boldsymbol{W^O} + \boldsymbol{Q_l}, \end{aligned} \tag{2}$$

with

$$\begin{aligned} head_i = & \text{ Attention } \left(Q = \boldsymbol{Q_l}\boldsymbol{W_i^Q}, K = \boldsymbol{X}\boldsymbol{W_i^K}, V = \boldsymbol{X}\boldsymbol{W_i^V}\right), \\ \text{Attention } (Q, K, V) = & \text{ softmax } (QK^T/\sqrt{d_h})V, \end{aligned} \tag{3}$$

where $\boldsymbol{Q_l} \in \mathbb{R}^{S \times h}$ is a learnable sequence of $S$ $h$-dimensional query vectors, the matrices $\boldsymbol{W}$ are learnable linear projections, $d_h$ is a scaling factor commonly set as the layer's hidden dimension, and $k$ is the number of attention heads of the layer. Both $S$ and $h$ are hyperparameters independent of the shape of the original sequence $\boldsymbol{X}$, and it is $S$ which defines the output sequence length of the MHA operation in SeqShort.

This MHA operation has a sorting effect: independent of the arrangement of the patch feature vectors in $\boldsymbol{X}$, the first row of $\boldsymbol{X_S}$ aggregates the instances that the first query vector in $\boldsymbol{Q_l}$ agrees with the most; the second row of $\boldsymbol{X_S}$ aggregates the instances that the second query vector in $\boldsymbol{Q_l}$ agrees with the most, and so on. This enables to incorporate positional information in our model based on a new interpretation: instead of thinking of the original arrangement of instances in the WSI 2D space, we consider the order of the rows of $\boldsymbol{X_S}$ as the available positional information possible to encode.

The resulting time complexity of the MHA operation performed by our input layer is $O(n)$ because of the fixed-size $\boldsymbol{Q_l}$, and since SeqShort is a single layer, the main compute load lies in the subsequent deeper transformer model in our pipeline. Although our method does not change the computational complexity of the MHSA layers of the transformer itself, by performing sequence reduction, the amount of FLOPs and memory it requires becomes constant with respect to the original number of WSI patches. The result is an overall considerable reduction of computational cost. Fig. 1c visualizes how the required FLOPs per forward pass scale better when using the SeqShort layer. For example, the average WSI in the IBC dataset comprises 7690 patches, which takes 734ms to be processed with a BERT-base encoder using our hardware. This time is reduced to 14ms when SeqShort is used as input layer.

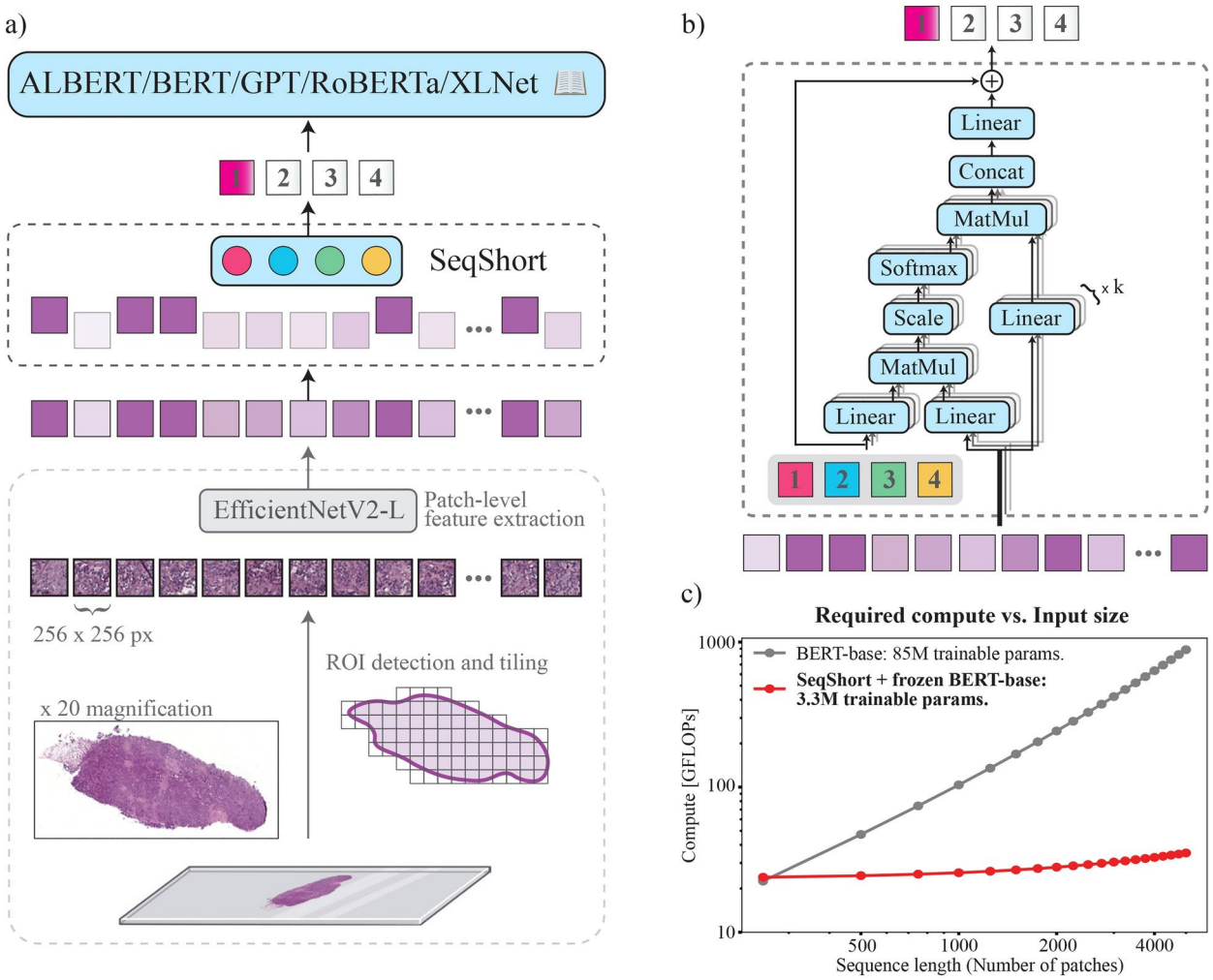



**Fig. 1**. Proposed method. (**a**) From bottom upwards: after a typical MIL pre-processing step (tiling, feature extraction), our SeqShort layer using a pre-defined number of query vectors (colored circles) summarises the long list of patches into a small, ordered sequence of feature vectors (colored squares) which are then classified by a deep transformer model that was pre-trained on an extensive text corpus. Different patches are in varying proportion part of the resulting feature vectors which is symbolically represented by their color intensity. (**b**) Detailed view of the SeqShort layer, where a set of learnable vectors (colored squares) query the relevant information in the WSI patches via a multi-head attention operation. (**c**) The computational cost of a forward pass of a deep transformer classifier is considerably reduced when our SeqShort layer is used (measured with the fvcore library by FAIR[1]).

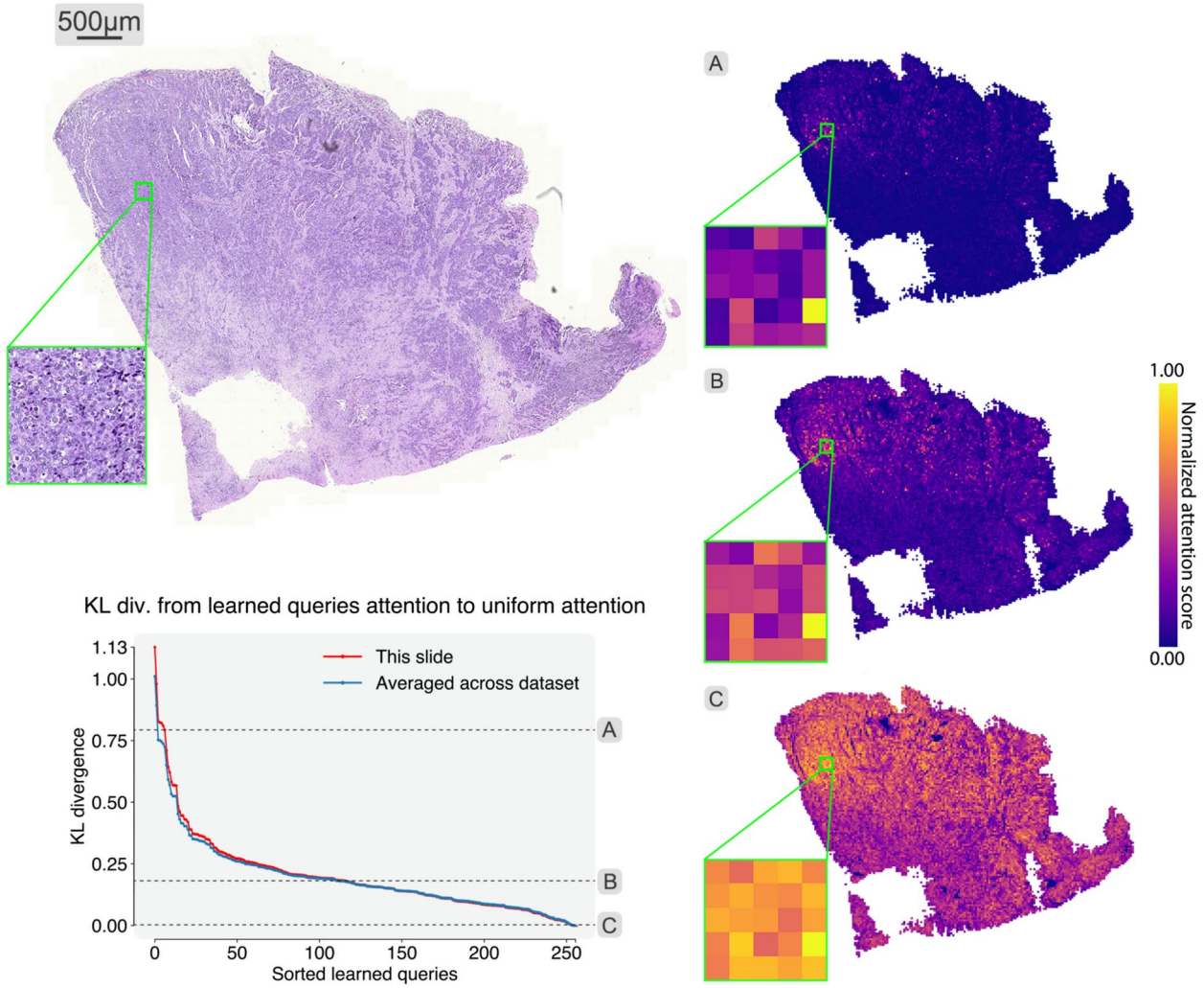



**Fig. 2**. WSI summarization. A WSI, and attention heatmaps (**A**–**C**) produced by three different query vectors in SeqShort are shown. Although different queries show attention distributed over a broader or narrower set of patches, the most important instances agree among the heatmaps. The bottom left plot shows the Kullback-Leibler divergence from the attention distributions of the learned queries to uniform attention, and the values that correspond to the heatmaps are indicated with dashed lines. Values are sorted for ease of visualization, higher values correspond to uneven distribution of attention among patches.

## Transformer models

In our experiments we find that a BERT-base encoder[6] pre-trained with the masked language modeling task of *Robustly optimized BERT pre-training Approach* (RoBERTa)[8] on a corpora of more than 160GB of uncompressed text comprised by BookCorpus[56], CC-News[57], OpenWebText[58] and Stories[59] yields the best results. We discard the vocabulary embeddings lookup table of RoBERTa-base as it is not needed for weakly supervised image classification.

## Complete pipeline

Our pipeline is illustrated in Fig. 1. As a pre-processing step, we extract non-overlapping tissue tiles of $256 \times 256$ pixels from each WSI. Tissue segmentation is done as in[60]. We use $\times 20$ and $\times 10$ magnification in different experiments. We generate the instance-level feature vectors using an EfficientNetV2-L[43] pre-trained on ImageNet[45].

The complete weakly supervised architecture that performs classification on the bag of instance vectors is composed of the SeqShort layer and a transformer language model. We set the vector embedding dimension of SeqShort to $h = 768$ (the hidden dimension of the used transformers), and $k = 4$ attention heads. For the lymph node classification task we set the output length of SeqShort to $S = 511$, and for the cancer subtyping tasks, to $S = 256$. A learnable [CLS] token is concatenated to the output of SeqShort, and added a sequence of learnable positional embeddings. The last hidden representation of [CLS] is the input of a multilayer perceptron (MLP) classification head. Altogether, our model comprises a total of 3.3M trainable parameters.

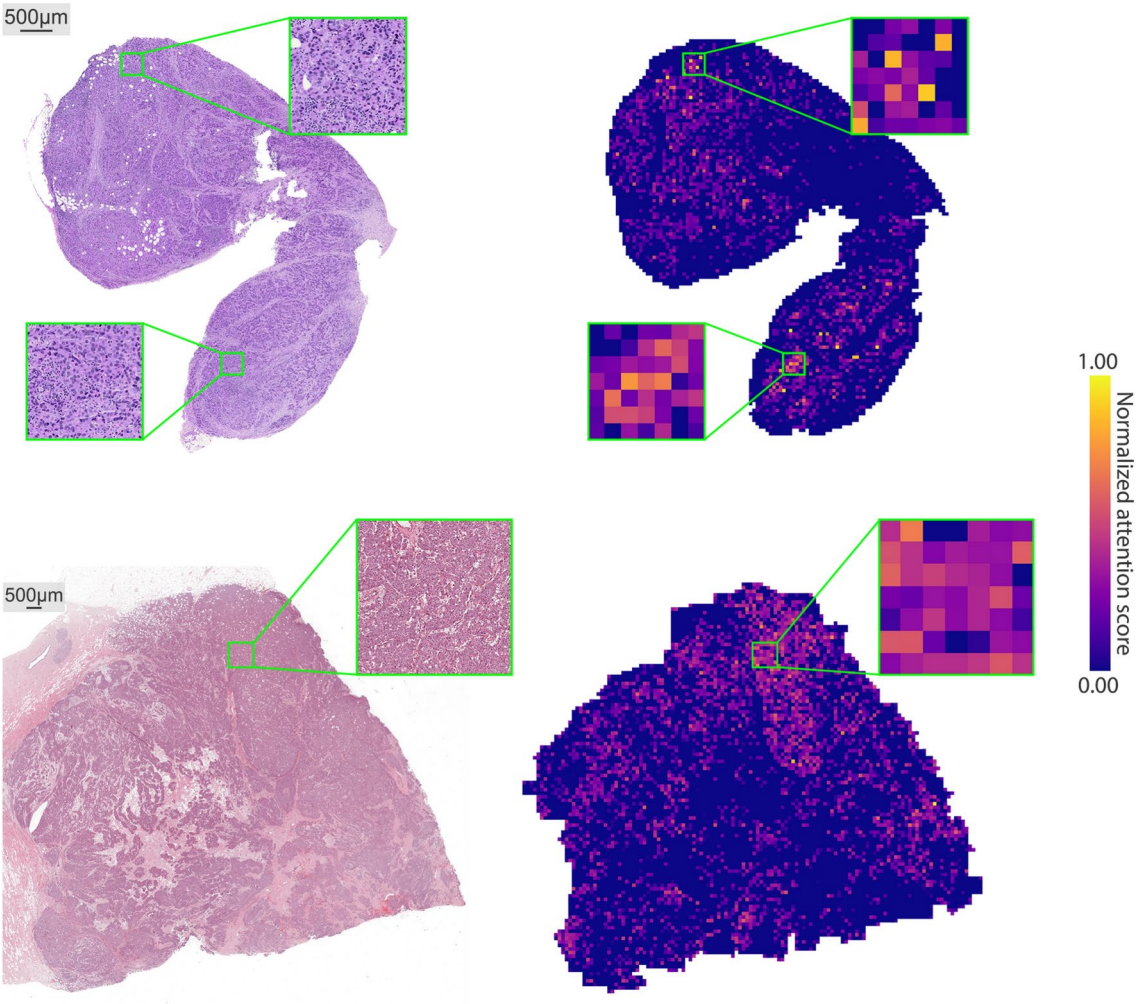



**Fig. 3**. Attention rollout heatmaps. Left: original WSIs. Right: their corresponding attention rollout heatmaps. Although the SeqShort layer reduces the number of feature vectors that the downstream transformer has to process, it is still possible to backtrack the contribution of each individual image patch to the classification decision using this most common transformer explainability method.

## Implementation and training

The method was implemented in Python, using PyTorch[61] as deep learning back-end. The pre-trained weights of EfficientNetV2-L and RoBERTa were downloaded from Torchvision[62] and HuggingFace[63], respectively. Training of our models was done with the aid of PyTorch-Lightning[64], on a single NVIDIA Tesla V100 GPU. The code of this project is available at https://github.com/bozeklab/lmagp/.

All our models were trained for 200 epochs. For the lymph node classification task the first 5 epochs were used as learning rate warm-up stage, followed by one cycle of a cosine schedule, with a maximum learning rate of $1 \times 10^{-4}$, and batch size of 16. For the cancer subtyping tasks, the warm-up stage lasted 10 epochs, followed by two cycles of a cosine schedule, with a maximum learning rate of $5 \times 10^{-5}$, and batch size of 32. Adam[65] was used as optimization algorithm.

## Datasets

*Lymph node metastases classification*

For this task we used the dataset provided by the CAMELYON16 grand challenge (https://camelyon16.grand-challenge.org/) which comprises 400 Hematoxylin and Eosin (H&E) stained WSIs of sentinel lymph nodes of breast cancer patients, scanned by 3DHISTECH and Hamamatsu scanners at $\times 40$ at the Radboud University Medical Center and the University Medical Center Utrecht, Netherlands. The grand challenge dataset is divided in a train set of 270 WSIs (160 normal slides, and 110 slides containing metastases), and a test set of 129 WSIs (80 normal slides, 49 slides containing metastases). In our experiments, we divided the provided train set in 90%/10% stratified splits for training and validation, respectively.

*Invasive breast carcinoma subtype classification*
We use a subset of 1,038 H&E stained WSIs from the TCGA-BRCA project within The Cancer Genome Atlas repository (https://www.cancer.gov/tcga). Out of the 1,038 slides, 889 were of patients with Invasive Ductal Carcinoma, and 149 were of patients with Invasive Lobular Carcinoma. We follow the study design in[40,44] and evaluate the models using stratified 10-fold cross-validation on patient level.

*Renal cell carcinoma subtype classification*
We use 918 H&E stained WSIs of Renal Cell Carcinoma cases from the TCGA repository. Out of these samples, 289 were of Chromophobe Cell Carcinoma patients, 118 were of Papillary Cell Carcinoma patients, and 498 were of Clear Cell Carcinoma patients, coming from the TCGA-KICH, TCGA-KIRP and TCGA-KIRC projects, respectively. We follow the same study design as in the IBC subtype classification task, and evaluate the models using stratified 10-fold cross-validation on patient level.

## Data availibility
Data used in this article comes from The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) and the CAMELYON16 Grand Challenge (https://camelyon16.grand-challenge.org/).

Received: 12 September 2024; Accepted: 24 January 2025
Published online: 15 February 2025

## References
1. FAIR (Last accessed: 12.09.2024) fvcore library. https://github.com/facebookresearch/fvcore/.
2. Vaswani, A. et al. Attention is all you need. *NeurIPS* **2017**, 30 (2017).
3. Ott, M. et al. Scaling neural machine translation. In *WMT 2018* 1–9 (Association for Computational Linguistics, 2018).
4. So, D., Le, Q., Liang, C. The evolved transformer. In: ICML 2019, PMLR, pp 5877–5886 (2019).
5. Dehghani, M., Gouws, S., Vinyals, O., et al. Universal transformers.https://doi.org/10.48550/ARXIV.1807.03819 (2018).
6. Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
7. Lan, Z., Chen, M., Goodman, S., et al. ALBERT: A lite BERT for self-supervised learning of language representations. In: ICLR (2020).
8. Liu, Y., Ott, M., Goyal, N., et al. Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019).
9. Radford, A., Narasimhan, K., Salimans, T., et al. Improving language understanding by generative pre-training. (2018).
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021).
11. Liu, Z., Lin, Y., Cao, Y., et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. IEEE, pp 9992–10002 (2021).
12. Zhai, X., Kolesnikov, A., Houlsby, N., et al. Scaling vision transformers. In: CVPR. IEEE, pp 1204–1213 (2022).
13. Zheng, S. et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR. Computer Vision Foundation* (ed. Zheng, S.) 6881–6890 (IEEE, 2021).
14. Ramesh, A., Pavlov, M., Goh, G., et al. Zero-shot text-to-image generation. In: Meila M, Zhang T (eds) ICML, PMLR, vol 139. PMLR, pp 8821–8831 (2021).
15. Saharia, C., Chan, W., Saxena, S., et al. Photorealistic text-to-image diffusion models with deep language understanding. CoRR abs/2205.11487 (2022).
16. Dietterich, T. G., Lathrop, R. H. & Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **89**(1–2), 31–71 (1997).
17. Maron, O. & Lozano-Pérez, T. A framework for multiple-instance learning. In *NeurIPS* (eds Jordan, M. I. et al.) 570–576 (The MIT Press, 1997).
18. Reisenbüchler, D. et al. Local attention graph-based transformer for multi-target genetic alteration prediction. In *MICCAI 2022, Part II, LNCS* Vol. 13432 (eds Wang, L., Dou, Q., Fletcher, P. T. et al.) 377–386 (Springer, 2022).
19. Lv, Z. et al. Joint region-attention and multi-scale transformer for microsatellite instability detection from whole slide images in gastrointestinal cancer. In *MICCAI 2022, Part II* Vol. 13432 (eds Wang, L., Dou, Q., Fletcher, P. T. et al.) 293–302 (Springer, 2022).
20. Shao, Z., Bian, H., Chen, Y., et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Ranzato M, Beygelzimer A, Dauphin YN, et al (eds) NeurIPS 2021, pp 2136–2147 (2021).
21. Li, H., Yang, F., Zhao, Y., et al. DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: de Bruijne M, Cattin PC, Cotin S, et al (eds) MICCAI 2021, Part VIII, LNCS, vol 12908. Springer, pp 206–216 (2021).
22. Zhao, Y. et al. SETMIL: spatial encoding transformer-based multiple instance learning for pathological image analysis. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022–25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, LNCS* Vol. 13432 (eds Wang, L., Dou, Q., Fletcher, P. T. et al.) 66–76 (Springer, 2022).
23. Shao, Z. et al. LNPL-MIL: learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023* (ed. Shao, Z.) 21438 (IEEE, 2023).
24. Tang, W. et al. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16–22, 2024* (ed. Tang, W.) 11343–11352 (IEEE, 2024).
25. Zheng, Y. et al. A graph-transformer for whole slide image classification. *IEEE Trans. Med. Imaging* **41**(11), 3003–3015 (2022).
26. Wang, S., Li, BZ., Khabsa, M., et al. Linformer: Self-attention with linear complexity. CoRR abs/2006.04768 (2020).
27. Dai, Z., Lai, G., Yang, Y., et al. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) NeurIPS 2020 (2020).
28. Clark, K., Khandelwal, U., Levy, O., et al. What does BERT look at? an analysis of bert's attention. In: Linzen T, Chrupala G, Belinkov Y, et al (eds) ACL 2019. Association for Computational Linguistics, pp 276–286 (2019).
29. Beltagy, I., Peters, ME., Cohan, A. Longformer: The long-document transformer. CoRR abs/2004.05150 (2020).
30. Zaheer, M., Guruganesh, G., Dubey, KA., et al. Big bird: Transformers for longer sequences. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) NeurIPS 2020 (2020).
31. Roy, A. et al. Efficient content-based sparse attention with routing transformers. *Trans. Assoc. Comput. Linguistics* **9**, 53–68 (2021).
32. Qiu, J. et al. Blockwise self-attention for long document understanding. In *EMNLP 2020, Findings of ACL, vol EMNLP 2020* (eds Cohn, T. et al.) 2555–2565 (Association for Computational Linguistics, 2020).
33. Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training verifiers to solve math word problems. CoRR abs/2110.14168 (2021).

34. Deletang, G., Ruoss, A., Duquenne, PA., et al. Language modeling is compression. In: The Twelfth International Conference on Learning Representations (2024).
35. Huang, W., Abbeel, P., Pathak, D., et al. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: Chaudhuri K, Jegelka S, Song L, et al (eds) ICML 2022, PMLR, vol 162. PMLR, pp 9118–9147 (2022).
36. Ilharco, G., Zellers, R., Farhadi, A., et al. Probing contextual language models for common ground with visual representations. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) NAACL-HLT 2021. Association for Computational Linguistics, pp 5367–5377 (2021).
37. Lu, K., Grover, A., Abbeel, P., et al. Frozen pretrained transformers as universal computation engines. In: AAAI-IAAI-EAAI 2022. AAAI Press, pp 7628–7636 (2022).
38. Abnar, S., Zuidema, W.H. Quantifying attention flow in transformers. In: Jurafsky D, Chai J, Schluter N, et al (eds) ACL 2020. Association for Computational Linguistics, pp 4190–4197 (2020).
39. Ilse, M., Tomczak, JM., Welling, M. (2018) Attention-based deep multiple instance learning. In: Dy JG, Krause A (eds) ICML 2018, PMLR, vol 80. PMLR, pp 2132–2141
40. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**(6), 555–570 (2021).
41. Li, B., Li, Y., Eliceiri, KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR 2021. Computer Vision Foundation / IEEE, pp 14318–14328 (2021).
42. Wagner, S. J. et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**(9), 1650-1661.e4 (2023).
43. Tan, M. & Le, Q. V. Efficientnetv2: Smaller models and faster training. In *ICML 2021* Vol. 139 (eds Meila, M. & Zhang, T.) 10096–10106 (PMLR, 2021).
44. Chen, RJ., Chen, C., Li, Y., et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR 2022. IEEE, pp 16123–16134 (2022).
45. Deng, J., Dong, W., Socher, R., et al. Imagenet: A large-scale hierarchical image database. In: CVPR 2009, Ieee, pp 248–255 (2009).
46. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559. https://doi.org/10.1016/j.media.2022.102559 (2022).
47. Yang, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS* **2019**, 32 (2019).
48. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019).
49. Dubey, A., Jauhri, A., Pandey, A., et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024).
50. Zhang, S., Roller, S., Goyal, N., et al. OPT: open pre-trained transformer language models. CoRR abs/2205.01068. https://doi.org/10.48550/ARXIV.2205.01068 (2022).
51. Nawrot, P. et al. Hierarchical transformers are more efficient language models. In *NAACL 2022* (eds Carpuat, M. et al.) 1559–1571 (Association for Computational Linguistics, 2022).
52. Liu, Y., Li, L., Zhang, B., et al. Matcr: Modality-aligned thought chain reasoning for multimodal task-oriented dialogue generation. In: El-Saddik A, Mei T, Cucchiara R, et al (eds) Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023. ACM, pp 5776–5785 (2023).
53. Zhang, H., Li, X. & Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6–10, 2023* (eds Feng, Y. & Lefever, E.) 543–553 (Association for Computational Linguistics, 2023).
54. Li, J., Li, D., Savarese, S., et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause A, Brunskill E, Cho K, et al (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, Proceedings of Machine Learning Research, vol 202. PMLR, pp 19730–19742 (2023).
55. Carion, N. et al. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020–16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* Vol. 12346 (eds Vedaldi, A., Bischof, H., Brox, T. et al.) 213–229 (Springer, 2020).
56. Zhu, Y., Kiros, R., Zemel, RS., et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: ICCV 2015. IEEE Computer Society, pp 19–27 (2015).
57. Mackenzie, JM., Benham, R., Petri, M., et al. Cc-news-en: A large english news corpus. In: d'Aquin M, Dietze S, Hauff C, et al (eds) CIKM 2020. ACM, pp 3077–3084 (2020).
58. Gokaslan, A., Cohen, V. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus (2019).
59. Trinh, TH., Le, QV. A simple method for commonsense reasoning. CoRR abs/1806.02847 (2018).
60. Graham, S. et al. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
61. Paszke, A., Gross, S., Massa, F., et al. Pytorch: An imperative style, high-performance deep learning library. In: Wallach HM, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp 8024–8035, https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (2019).
62. Marcel, S., Rodriguez, Y. Torchvision the machine-vision package of torch. In: Bimbo AD, Chang S, Smeulders AWM (eds) Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010. ACM, pp 1485–1488. https://doi.org/10.1145/1873951.1874254 (2010).
63. Wolf, T., Debut, L., Sanh, V., et al. Huggingface's transformers: State-of-the-art natural language processing. CoRR abs/1910.03771. http://arxiv.org/abs/1910.03771,1910.03771 (2019).
64. Falcon, W. The PyTorch Lightning team. PyTorch Lightning.https://doi.org/10.5281/zenodo.3828935, https://github.com/Lightning-AI/lightning (2019).
65. Kingma, DP., Ba, J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, http://arxiv.org/abs/1412.6980 (2015).


## Acknowledgements

We thank Noémie C. Moreau and Philipp Antczak for the fruitful discussions and feedback. Both K.B. and J.I.P. were hosted by the Center for Molecular Medicine Cologne throughout this research. K.B. and J.I.P. were supported by the BMBF program Junior Group Consortia in Systems Medicine (01ZX1917B), BMBF program for Female Junior Researchers in Artificial Intelligence (01IS20054) and the NRW return program.

## Author contributions

J.I.P. and K.B. wrote the manuscript. J.I.P developed the software, conducted experiments, and prepared the article figures.

## Funding
Open Access funding enabled and organized by Projekt DEAL.

## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-88139-5.

**Correspondence** and requests for materials should be addressed to J.I.P.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 6

# Fine-tuning a Multiple Instance Learning Feature Extractor with Masked Context Modelling and Knowledge Distillation

While the previous chapters dealt with the classification task in a MIL pipeline, the final work in this dissertation focuses on the feature extraction step. Feature extractor models are pre-trained to produce meaningful representations of their image input, and a MIL model is trained to generate useful representations of their task-specific, multiple-instance input. This work proposes to link these two aspects by considering a fundamental visual characteristic of WSIs: the image patches derived from a tissue scan are not isolated entities but are instead highly correlated with their surrounding neighborhoods.

# Fine-tuning a Multiple Instance Learning Feature Extractor with Masked Context Modelling and Knowledge Distillation

Juan I. Pisula[1,2] and Katarzyna Bozek[1,2,3]

[1] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany
[2] Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany
[3] Cologne Excellence Cluster on Cellular Stress Responses in Aging- Associated Diseases (CECAD), University of Cologne, Germany
juan.pisula@uk-koeln.de, k.bozek@uni-koeln.de

**Abstract.** The first step in Multiple Instance Learning (MIL) algorithms for Whole Slide Image (WSI) classification consists of tiling the input image into smaller patches and computing their feature vectors produced by a pre-trained feature extractor model. Feature extractor models that were pre-trained with supervision on ImageNet have proven to transfer well to this domain, however, this pre-training task does not take into account that visual information in neighboring patches is highly correlated. Based on this observation, we propose to increase downstream MIL classification by fine-tuning the feature extractor model using *Masked Context Modelling with Knowledge Distillation*. In this task, the feature extractor model is fine-tuned by predicting masked patches in a bigger context window. Since reconstructing the input image would require a powerful image generation model, and our goal is not to generate realistically looking image patches, we predict instead the feature vectors produced by a larger teacher network. A single epoch of the proposed task suffices to increase the downstream performance of the feature-extractor model when used in a MIL scenario, even capable of outperforming the downstream performance of the teacher model, while being considerably smaller and requiring a fraction of its compute.

**Keywords:** Multiple Instance Learning · Masked Context Modelling · Knowledge Distillation

## 1 Introduction

In Digital Pathology, specimen slides are digitised into high resolution Whole Slide Images (WSIs) of several gigapixels. This has led to the popularisation of Multiple Instance Learning (MIL) [12, 33] algorithms for automatic WSI classification tasks. In these algorithms, the typical pre-processing step involves tesselating the WSI into smaller image patches, or instances, and computing their

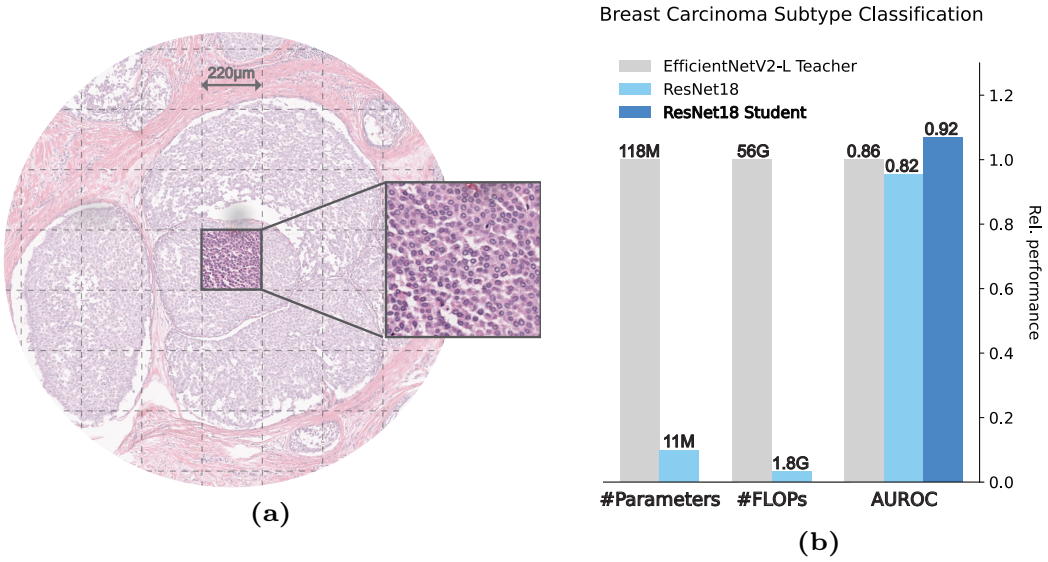



**Fig. 1:** a) A cutout of a Breast Carcinoma HE slide, where a highlighted image patch shows a cluster of cells. When inspecting its neighborhood, it is seen that this cluster is not an isolated pattern, but part of a mammary lobe. *Masked Context Modelling with Knowledge Distillation* aims to improve downstream performance by including context information in the feature extraction step. b) Comparison (number of parameters, FLOPs per forward pass, downstream MIL classification task AUROC) of ImageNet pre-trained feature extraction models: EfficientNetV2-L, ResNet18, and ResNet18 fine-tuned with our method using the EfficientNetV2-L as teacher. CLAM was used as MIL classification model, and performance is visualized relative to the EfficientNetV2-L model.

feature vector representations using a pre-trained feature extractor neural network. To compute patch representations a popular choice is to use models pre-trained with supervision on ImageNet [37]. Following this step, a MIL model is trained to produce a slide-level prediction using the complete set of instances of the WSI.

As noted in [3], the common characteristic in image data is that neighboring pixels are highly correlated, and this fact extends to neighboring image patches in histology slides. A cutout of a Breast Carcinoma Hematoxylin and Eosin (HE) slide is shown in Fig. 1a), with a highlighted image patch of 220$\mu$m or 224 pixels side length at 10$\times$ magnification level (approximately 1$\mu$m/pixel), showing a cluster of cells. HE image patches of these dimensions can capture fine-grained histology features like individual cell nucleus morphology, cell conglomerates, and small functional structures such as glands and blood vessels. These patterns tend to extend for several patches, and coarse-grained features, such as distribution and density of cell populations, arrangement of functional structures, tissue interfaces, overall tissue morphology and architecture, are revealed when examining a broader context. In a MIL pipeline, although a pre-trained feature extractor model can produce useful instance-level representations, it is the downstream MIL classifier that should detect and make sense out of the histological patterns that take more image area than a single patch.

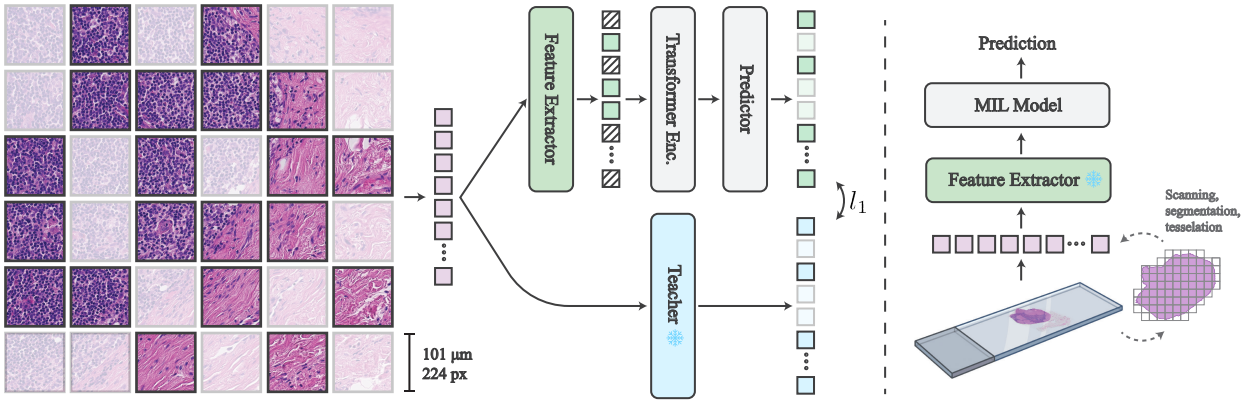



**Fig. 2:** Proposed pipeline. During the feature extractor fine-tuning stage (left), a pre-trained feature extractor model is fed with image patches coming from a larger context window. A random subset of the patches' feature vector representations is masked, and a Transformer encoder with a predictor network is used to predict the masked instances' feature vector representations produced by a frozen teacher network, minimizing an $l_1$ loss. For the downstream task training stage (right), the Transformer and the predictor networks are discarded, and the fine-tuned feature extractor can be used in any Multiple Instance Learning pipeline.

In this work we propose to improve downstream MIL classification by fine-tuning the feature extractor model, such that representations of neighboring patches are predictive of one another. Drawing inspiration from reconstruction-based algorithms in the Self-Supervised Learning (SSL) literature [11, 19, 35, 36, 46–48], we propose the Masked Context Modelling (MCM) task: individual image patches from a bigger context window are masked, and the feature extractor model is fine-tuned to predict the missing patches based on the visible ones.

As described above, the MCM task is challenging as there are multiple biologically plausible ways to fill the missing image patches in the masked context. Additionally, synthesizing such image patches would require an image generation model, such as a Variational Autoencoder [27], a Generative Adversarial Network [15], or a Diffusion Model [23, 41]. This image generation task is secondary, as our goal is not to produce realistically looking image patches but to learn useful representations. We propose therefore to predict instead of images, the feature vectors of a larger, pre-trained, teacher network and using them as a proxy of the visual information of the masked images.

The contributions of this work can be summarised as follows:

- We introduce the *Masked Context Modelling with Knowledge Distillation (MCM+KD)* task to fine-tune the feature extractor model used in a MIL pipeline by using a larger pre-trained model as teacher.
- This task can be trained even for a single epoch, between the feature extractor pre-training stage and the final training of the MIL model, improving downstream performance in two Cancer Subtype Classification tasks and a Lymph Node Metastases Detection task.
- We show that the fine-tuned student does not learn to copy the teacher's output. Furthermore, the student model can result in higher downstream

performance than the teacher, while having fewer parameters and needing less computation to process input images (Fig. 1b).

## 2 Related Work

### 2.1 Pre-training and Fine-tuning

It is empirically proven that transfer learning from ImageNet can improve classification performance and speed up convergence in medical image analysis applications [1, 16, 20]. In MIL algorithms for Digital Pathology, frozen ImageNet pre-trained models are commonly used in the feature extraction step [4,28,30,39]. This is partially due to the fact that in a MIL problem, the lack of instance-level labels prevents the fine-tuning of a feature extractor model with supervision.

A different line of research investigates how the domain shift between natural images and medical images could be alleviated by algorithms that require no labels. For example, [2] proposes to do SSL with in-domain medical data on ImageNet pre-trained models, prior to supervised fine-tuning. We refer to [18] for a comprehensive survey on Domain Adaptation in the medical image field, including unsupervised methods.

### 2.2 Reconstruction-based Algorithms

Training models without supervision – by masking a part of their input and teaching them to reconstruct it – has achieved great success both in Computer Vision (CV) and Natural Language Processing (NLP). Pioneering works in CV include Denoising Autoencoders [46] and Context Encoders [35]. In the former, masking is applied as noise corruption, where in the latter, a region of the input image is explicitly zeroed out.

With the advent of the Transformer architecture [45], Masked Language Modelling (MLM) [11] and Causal Language Modelling [36] became the dominant SSL algorithms in NLP. MLM has inspired a plethora of SSL works in CV, yielding the Masked Image Modelling (MIM) family of algorithms [19, 47, 48], although these methods are restricted to Vision Transformer (ViT) architectures [13].

### 2.3 Knowledge Distillation

Knowledge Distillation (KD) consists of training a student neural network to predict outputs produced by a teacher model. It was originally proposed by [22] as a model compression method, by using the prediction outputs of a larger model as soft label targets to train a smaller model. Several variants have been proposed, including the prediction of the teacher's intermediate activations, confidence maps for pose estimation [49], or the use of a special token for distillation in Transformer models [44]. Notably, these model compression techniques have been applied in non-MIL image classification tasks in Digital Pathology [6, 25, 26, 32].

Knowledge Distillation has become increasingly popular in the SSL literature with Self-Distillation (SD) algorithms, where the pre-trained teacher network is

replaced by an Exponential Moving Average (EMA) of the student [5, 9, 17, 50]. Recently in Digital Pathology, [43] has proposed the use of an additional SD-based loss for training supervised MIL classifiers.

## 3 Methodology

We now introduce our *Masked Context Modelling with Knowledge Distillation* task, designed to fine-tune the feature extractor network from the pre-processing step of a MIL pipeline. This fine-tuning step is performed before training of the WSI classification MIL model, and is independent of this model's architecture. The proposed pipeline is illustrated in Fig. 2, and described below.

**Context Window**. Given a MIL pipeline where a WSI is processed as a set of square image patches of side $p$ at a specified magnification objective, we crop large square patches of side $P = s \cdot p$ at the same magnification objective. The large image patches act as context windows, and are tessellated into $s \cdot s$ square patches of side $p$.

**Student and Teacher**. The feature extractor network $f$ to be fine-tuned acts as a student network. $f$ converts the image patches of a context window into the sequence of feature vectors $\mathbf{x} \in \mathbb{R}^{s \cdot s \times d_f}$, where $d_f$ is the latent dimension of the model. Similarly, the same image patches are converted into the sequence of feature vectors $\mathbf{y} \in \mathbb{R}^{s \cdot s \times d_t}$ by the network $t$, a frozen and pre-trained teacher model with latent dimension $d_t$. Feature vectors from the sequence $\mathbf{y}$ are used later as prediction targets.

**Masked Context Modelling**. Following the masking strategy of [19, 48], a random subset of instances from $\mathbf{x}$ is sampled using a uniform distribution, and each of the sampled instances is replaced by a learnable mask token. We add learnable positional embeddings to this sequence and feed it to a Transformer encoder model, which outputs the latent sequence representation of the masked context window, $\mathbf{x}_L \in \mathbb{R}^{s \cdot s \times d_f}$.

**Predictor Network and Objective**. A predictor network predicts the feature vector sequence $\mathbf{y}$ produced by the teacher, from the latent representation of the masked context, $\mathbf{x}_L$. The whole pipeline is trained to minimize the $l_1$ loss as in SimMIM [48]:

$$L = \frac{1}{\Omega(\mathbf{y}_M)} \|\mathbf{y}_M - \mathbf{y}'_M\|_1, \tag{1}$$

where $\mathbf{y}' \in \mathbb{R}^{s \cdot s \times d_t}$ is the output of the predictor network, the subscript $M$ denotes the set of feature vectors that correspond to masked instances, and $\Omega(\cdot)$ is the number of elements.

Once $f$ is fine-tuned, the Transformer encoder and the predictor network are discarded. Then $f$ can be used as feature extractor network in the pre-processing step of the downstream MIL task.

# 4 Experiments

## 4.1 Datasets and Pre-processing

We evaluate our method in the tasks of Breast Carcinoma Subtype Classification (BCSC), Lung Carcinoma Subtype Classification (LCSC), and Lymph Node Metastases Detection.

**Breast Carcinoma Subtype Classification**. This dataset consists of 500 HE-stained WSIs from the TCGA-BRCA project within The Cancer Genome Atlas repository (https://www.cancer.gov/tcga). Out of the 500 slides, 351 come from Invasive Ductal Carcinoma cases, and 149 come from Invasive Lobular Carcinoma cases. Evaluation was done in a train-val-test fashion, using patient-level stratified data splits of 80%-10%-10%.

**Lung Carcinoma Subtype Classification**. This dataset consists of 500 HE-stained WSIs from the TCGA-LUAD and TCGA-LUSC projects within The Cancer Genome Atlas repository. Out of the 500 slides, 250 come from Lung Adenocarcinoma cases, and 250 come from Lung Squamous Cell Carcinoma cases. Same evaluation protocol as in the BCSC task was used.

**Lymph Node Mestastases Detection**. This dataset is provided by the CAMELYON16 Grand Challenge (https://camelyon16.grand-challenge.org/). The training dataset comprises 270 HE-stained WSIs of sentinel lymph nodes (160 normal slides, and 110 slides containing metastases), and the test set 129 WSIs (80 normal slides, 49 containing metastases). The slides were scanned by 3DHISTECH and Hamamatsu scanners at $\times 40$ magnification objective at the Radboud University Medical Center and the University Medical Center Utrecht, Netherlands. In our experiments, we divided the provided train set in 90%-10% stratified data splits for training and validation.

During pre-processing, the slides were tiled into large crops of $1792 \times 1792$ pixels at a single magnification objective ($\times 20$ for the LNMD task and $\times 10$ for the BCSC and LCSC tasks), and images with less than 60% of tissue were removed. These crops acted as context windows for the MCM+KD task, and were subsequently subdivided into image patches of $224 \times 224$ pixels. In the downstream classification task, MIL models were trained considering all the $224 \times 224$ image patches of a single slide as an individual sample.

## 4.2 Experimental Setup

Throughout all the experiments we use an ImageNet pre-trained ResNet18 [38] as MIL feature extractor model and ImageNet pre-trained EfficientNetV2-L [42] as teacher model in the MCM+KD task. We take the activations before the last classification layer as feature vectors in both networks.

The Transformer encoder was parametrized with input size of 512 (the same as the ResNet18 feature vectors), 8 Multi-head Attention layers of 4 attention heads, and MLPs with hidden dimension of 3072. The predictor model is a 2-layer MLP with 1280 hidden units and output units (the same as the EfficientNetV2-L feature vectors), with a GELU [21] hidden activation layer.

We used MCM+KD for a single epoch to fine-tune the pre-trained ResNet18 in all experiments. AdamW optimizer [29] and batch size of 8 context windows were used. The learning rates were searched individually for each dataset, to maximize downstream CLAM [30] classification, yielding the learning rates of 0.0001, 0.001, 0.0005 for the LNMD, BCSC, and LCSC tasks, respectively, and for the other optimizer hyperparameters we kept PyTorch default values. The percentage of masked patches in a context was selected in a similar manner, masking 60% of the patches of a context window in the LNMD task, and 50% of the patches in a context window in the BCSC and LCSC tasks.

Downstream MIL classifiers were trained for 100 epochs, using a batch size of 1 WSI per batch, and accumulating gradients for 8 optimization steps. The models were optimized with AdamW, learning rate was searched for each dataset and classifier independently, and model selection was done based on validation loss. We report the macro-averaged AUROC as main performance metric.

The method was implemented in Python, using PyTorch [34] and PyTorch-Lightning [14]. The ImageNet pre-trained weights of the feature extractor models were downloaded from Torchvision [31]. Implementation of SSL methods was taken from the solo-learn library [10]. Training was done on a single NVIDIA Ampere A100 GPU. The code of this project is available at `https://github.com/bozeklab/mcm`.

### 4.3 WSI Classification Results

Our main experiment is the comparison of downstream MIL classification performance of a baseline ImageNet pre-trained ResNet18 used as feature-extractor model, against the same feature-extractor model, fine-tuned with MCM+KD for a single epoch. For the comparison, we include the downstream performance of MIL models that used the pre-trained EfficientNetV2-L teacher as feature extractor model.

We compare downstream classification of three popular MIL algorithms: Attention-Based Deep Multiple Instance Learning (ABDMIL) [24], which popularised the use of the attention mechanism in MIL; CLAM [30], an improvement over ABDMIL that incorporates a clustering loss in the latent space of instances; and TransMIL [39], a BERT-like Transformer encoder designed for Digital Pathology MIL tasks.

The results are shown in Tables 1-3. Our method increases the downstream performance of the ResNet18 feature extractor model across all performed comparisons, even surpassing the EfficientNetV2-L teacher model in some scenarios.

### 4.4 Self-Distillation Scenario

As suggested in [2], a possible solution to the domain shift problem when using ImageNet pre-trained models in medical image analysis is the use of SSL algorithms with in-domain data. In this experiment we use self-distillation where the pre-trained teacher model is replaced by an Exponential Moving Average (EMA) of the student network.

**Table 1:** Breast Carcinoma Subtype Classification results.

| Feature extraction | ABDMIL | CLAM | TransMIL |
|---|---|---|---|
| ResNet18 | 0.75 | 0.82 | 0.85 |
| ResNet18 (MCM+KD) | **0.76** | **0.92** | **0.89** |
| EfficientNetV2-L (teacher) | 0.72 | 0.86 | 0.94 |

**Table 2:** Lung Carcinoma Subtype Classification results.

| Feature extraction | ABDMIL | CLAM | TransMIL |
|---|---|---|---|
| ResNet18 | 0.83 | 0.86 | 0.68 |
| ResNet18 (MCM+KD) | **0.91** | **0.90** | **0.77** |
| EfficientNetV2-L (teacher) | 0.82 | 0.91 | 0.82 |

We compare this Self-Distillation version of the MCM task (denoted *MCM+SD*) against other SSL algorithms that make use of an EMA teacher: BYOL [17], DINO [5], MOCOv3 [9], and ReSSL [50]. We include SimCLR [7] and SimSiam [8] for comparison as well, although these methods are not from the Self-Distillation family.

This experiment was evaluated using CLAM downstream classification, and hyperparameters were kept the same as in the previous experiment. For the Self-Distillations algorithms, including MCM+SD, we use an EMA $\tau$ of 0.999.

The results in Table 4 show that our MCM+SD method was the only one improving the results over the baseline pre-trained ResNet18 for the BCSC and LCSC tasks. However, none of the tested methods improved LNMD classification. Although the compared methods are well established for self-supervised pre-training, our results highlight the role of the pre-trained teacher for fine-tuning a feature extractor model with the limited compute budget of a single epoch.

### 4.5 Ablation Study

Here we perform an ablation study to analyze individual components of MCM+KD (Table 5). In our experiment denoted as *MCM*, we fine-tune the feature extractor model by predicting the pixels of the masked image patches from the context window, omitting the knowledge distillation task. We also investigate the need of the MCM task: we omit the masking of patches and the Transformer encoder, and use a simple Knowledge Distillation (KD) approach instead. The KD consists of directly predicting the feature vectors of the teacher model.

Finally, we do an experiment (denoted by *CM + KD*) where the context modelling Transformer, predictor network, and teacher network are kept, but the patches are not masked. This test was performed to verify the need of such design choice. In contrast to an image reconstruction task, in the knowledge

**Table 3:** Lymph Node Metastases Detection results.

| Feature extraction | ABDMIL | CLAM | TransMIL |
|---|---|---|---|
| ResNet18 | 0.58 | 0.79 | 0.69 |
| ResNet18 (MCM+KD) | **0.69** | **0.84** | **0.76** |
| EfficientNetV2-L (teacher) | 0.76 | 0.75 | 0.76 |

**Table 4:** Comparison of downstream MIL classification using CLAM when doing pre-processing fine-tuning with different SSL methods, and a Self-Distillation version of the Masked Context Modelling task.

| Method | BCSC | LCSC | LNMD |
|---|---|---|---|
| Baseline | 0.82 | 0.86 | **0.79** |
| BYOL | 0.70 | 0.66 | 0.62 |
| DINO | 0.50 | 0.54 | 0.60 |
| MOCOv3 | 0.82 | 0.82 | 0.67 |
| ReSSL | 0.62 | 0.71 | 0.50 |
| SimCLR | 0.79 | 0.84 | 0.65 |
| SimSiam | 0.51 | 0.47 | 0.61 |
| MCM+SD | **0.88** | **0.89** | 0.65 |
| MCM+KD | 0.92 | 0.90 | 0.84 |

distillation scenario our model is not at risk of learning the identity function or memorizing the input data.

The ablation study was done with a CLAM downstream classifier, keeping the previous hyperparameters. For the MCM experiment, the MLP predictor was replaced by a single linear layer, as an MLP as described in 4.2 with $224 \times 224 \times 3$ hidden units is infeasible. Our results show that only our complete pipeline can improve the downstream CLAM performance over the baseline pre-trained ResNet18 feature extractor.

**Table 5:** Ablation study results.

| Model | LNMD | LCSC |
|---|---|---|
| Baseline | 0.79 | 0.86 |
| MCM | 0.42 | 0.72 |
| KD | 0.76 | 0.86 |
| CM + KD | 0.56 | 0.82 |
| MCM + KD | **0.84** | **0.90** |

### 4.6 Qualitative Results

The aim of this experiment is to visually evaluate the effect that our method has on the feature vectors of the ResNet18 feature extractor, independent of the downstream MIL classifier. We take slides from the CAMELYON16 test split which contain annotated metastases regions. We select a patch within a metastasis region and compute the cosine similarity of its feature vector and the feature vectors of the rest of the patches in the slide. We show these cosine similarity values as heatmap visualizations. We do this procedure with the feature vectors obtained by the baseline ResNet18, the ResNet18 fine-tuned with MCM+KD, and the EfficientNetV2-L teacher.

Fig. 3 shows heatmaps of a WSI with metastases. These heatmaps depict that our method decreases the similarity between feature vectors of metastases and feature vectors of normal tissue. The better visual discrimination between biologically different tissue regions in the heatmaps of the baseline and our fine-tuned ResNet18 models is consistent with the performance increase in the downstream MIL tissue classification task. The difference in the heatmaps of the teacher and the student models is consistent with the results in Sec. 4.5 showing that downstream performance is not achieved by copying the teacher's output.

## 5 Discussion and Conclusion

In this paper we demonstrate how downstream MIL classification can be improved by fine-tuning the feature extractor model using our proposed Masked Context Modelling with Knowledge Distillation task. This algorithm can be trained for a single epoch and is agnostic of the MIL model used in the downstream task. All our experiments were done using an EfficientNetV2-L teacher to fine-tune a ResNet18 feature extractor, with both networks being pre-trained with supervision on ImageNet.

Our results show that across the datasets and MIL models, our method improves the downstream performance. It is worth noting that although the teacher model is considerably bigger, computationally more expensive, and outperforms the ResNet18 in the original ImageNet classification pre-training task, the feature vectors it produces do not always result in better downstream performance. Nevertheless, the teacher model is still effective for distilling its knowledge to the student.

Using the same experimental setup, we fine-tuned the same ResNet18 with different SSL algorithms. These training scenarios did not however improve the results over the baseline ImageNet-pre-trained ResNet18. Our self-distillation version of the MCM task was not successful in the LNMD dataset, but it did improve the results in the BCSC and LCSC tasks, although not matching the results of MCM+KD with an EfficientNetV2-L teacher. This experiment suggests that even though SSL algorithms and self-distillation are effective in a full pre-training scenario, knowledge distillation from a larger pre-trained teacher is very useful in a single epoch of fine-tuning.

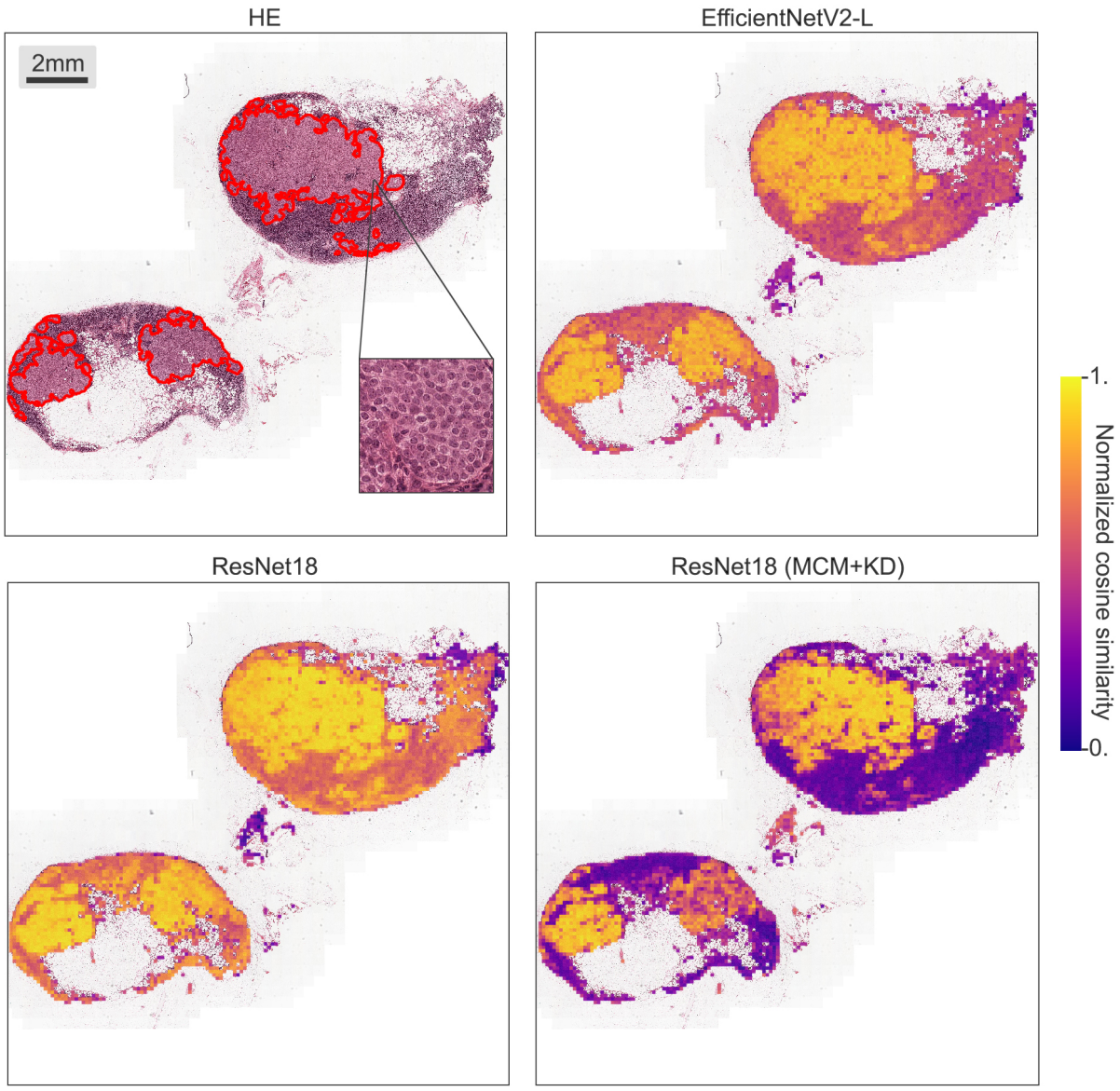



**Fig. 3:** Heatmap visualizations of the cosine similarity between the feature vector of a patch from a metastasis region and the rest of the feature vectors of the slide.

In our ablation study, we show that the individual components of our method do not work in a standalone setting, and the success of our method comes from the proper combination of these algorithms and principles. Our results show that the performance of our method is not explained solely by knowledge distillation, and we verify this by visual examination of cosine similarity heatmaps. The prediction of masked instances plays an essential role in our algorithm, and we leave exploring more sophisticated masking policies, such as adversarial masking [40], for future work.

Altogether, we indicate the importance of context in achieving the best representations of WSI patches. Our masking and KD approach allows to efficiently encode this context and to make use of it in the further WSI analysis tasks.

**Acknowledgements.** Both K.B. and J.I.P. were hosted by the Center for Molecular Medicine Cologne throughout this research. K.B. and J.I.P. were supported by the BMBF program Junior Group Consortia in Systems Medicine (01ZX1917B) and BMBF program for Female Junior Researchers in Artificial Intelligence (01IS20054).

## References

1. Alzubaidi, L., Fadhel, M.A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., R. Oleiwi, S.: Towards a better understanding of transfer learning for medical imaging: a case study. Applied Sciences **10**(13), 4523 (2020)
2. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3478–3488 (2021)
3. Burt, P.J., Adelson, E.H.: The laplacian pyramid as a compact image code. In: Readings in computer vision, pp. 671–679. Elsevier (1987)
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640. IEEE (2021)
6. Chaudhury, S., Shelke, N., Sau, K., Prasanalakshmi, B., Shabaz, M.: A novel approach to classifying breast cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization. Computational and Mathematical Methods in Medicine **2021**, 1–11 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 15750–15758. Computer Vision Foundation / IEEE (2021)
9. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9620–9629. IEEE (2021)
10. Da Costa, V.G.T., Fini, E., Nabi, M., Sebe, N., Ricci, E.: solo-learn: A library of self-supervised methods for visual representation learning. Journal of Machine Learning Research **23**(56), 1–6 (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1-2), 31–71 (1997)

13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Falcon, W., The PyTorch Lightning team: PyTorch Lightning (Mar 2019). `https://doi.org/10.5281/zenodo.3828935`, `https://github.com/Lightning-AI/lightning`
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
16. Graziani, M., Andrearczyk, V., Müller, H.: Visualizing and interpreting feature reuse of pretrained cnns for histopathology. In: MVIP 2019: Irish Machine Vision and Image Processing Conference Proceedings. Irish Pattern Recognition and Classification Society (2019)
17. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - A new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
18. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 15979–15988. IEEE (2022)
20. Heker, M., Greenspan, H.: Joint liver lesion segmentation and classification via transfer learning. arXiv preprint arXiv:2004.12352 (2020)
21. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J.G., Krause, A. (eds.) ICML 2018. PMLR, vol. 80, pp. 2132–2141. PMLR (2018)
25. Javed, S., Mahmood, A., Qaiser, T., Werghi, N.: Knowledge distillation in histology landscape by multi-layer features supervision. IEEE J. Biomed. Health Informatics **27**(4), 2037–2046 (2023)
26. Ke, J., Shen, Y., Wright, J.D., Jing, N., Liang, X., Shen, D.: Identifying patch-level msi from histological images of colorectal cancer by a knowledge distillation model. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1043–1046. IEEE (2020)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR 2021. pp. 14318–14328. Computer Vision Foundation / IEEE (2021)

29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2017)
30. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
31. Marcel, S., Rodriguez, Y.: Torchvision the machine-vision package of torch. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1485–1488 (2010)
32. Marini, N., Otálora, S., Müller, H., Atzori, M.: Semi-supervised learning with a teacher-student paradigm for histopathology classification: a resource to face data heterogeneity and lack of local annotations. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. pp. 105–119. Springer (2021)
33. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) NeurIPS. pp. 570–576. The MIT Press (1997)
34. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
35. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
36. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
38. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN computer science **2**(3), 160 (2021)
39. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) NeurIPS 2021. pp. 2136–2147 (2021)
40. Shi, Y., Siddharth, N., Torr, P., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: International Conference on Machine Learning. pp. 20026–20040. PMLR (2022)
41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
42. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
43. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4078–4087 (2023)
44. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357. PMLR (2021)

45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS 2017 **30** (2017)
46. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)
47. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 19175–19186. IEEE (2023)
48. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: a simple framework for masked image modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 9643–9653. IEEE (2022)
49. Zhang, F., Zhu, X., Ye, M.: Fast human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
50. Zheng, M., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., Xu, C.: Ressl: Relational self-supervised learning with weak augmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 2543–2555 (2021)

Chapter 7

# Discussion

Each chapter of this cumulative thesis includes its own discussion in the context of its specific study. Therefore, this section does not seek to reiterate previously covered points but instead focuses on commenting on additional dimensions of this research.

## 7.1 Chapter 2: Predicting HER2 overexpression from IHC-stained TMAs

As mentioned, this work represents the first project carried out during my doctoral studies, serving as both an introduction to the discipline and a foundational step in my education. It shares common characteristics with real-world, medical machine-learning projects. For instance, the class distribution of the data is highly imbalanced, with only 5% of the cases belonging to the positive class that overexpresses HER2. Consequently, it is easy to (mis)train a classifier to predict only the negative class if an improper metric is optimized, resulting in a high medical cost, as true positive therapy candidates might go undetected. In this context, having a rationale behind the model's decisions is crucial, as explainability and trustworthiness are essential in medical applications. Additionally, the dataset presents geographical biases, mainly manifested as batch effects in the appearance of images from different cohorts, potentially due to differences in tissue processing, and sample aging.

Certain aspects of this work, however, simplify the task compared to the other study cases presented in this dissertation. First, the images analyzed here are tissue cores with a side length of 5468 pixels — smaller than the WSIs utilized in the subsequent chapters. As a result, the MIL pipeline could be trained end-to-end, performing background removal and feature extraction "on the fly" instead of relying on an offline pre-processing stage. This streamlined the development of the algorithm and allowed the feature extractor to be trained under the supervision of the classification signal, learning image features optimized for this specific task.

Second, the task explored here is one that histopathologists routinely perform by visually inspecting IHC-stained TMAs. Unlike the therapy response tasks studied later, we know what to look for in these images, and there are experts capable of solving this task through visual examination. This fact is reflected in the study's results in multiple ways. The task can be solved with high performance metrics, as all trained machine learning models achieved good results (with varying degrees of success). Notably, the highest error rate occurred in images with an IHC score of 2 —equivocal cases that require an in-situ hybridization test for resolution, as they cannot be classified solely through visual inspection. Furthermore, the t-SNE plot in Fig. 2 suggests that the

data distribution lies along a one-dimensional manifold despite being embedded in a two-dimensional space, and visual inspection of the image patches reveals that the most prominent distinguishing feature is the IHC staining, essentially forming an IHC stain spectrum.

One challenge that first emerged during the development of this work was the external cohort evaluation of the model. Inter-center variability is a common issue in the quantitative analysis of histopathology images, stemming from differences in tissue processing steps such as staining duration, reagent concentrations, or sample storage conditions. Moreover, staining quality can deteriorate as samples age. In the CSCC study case, this issue is mitigated by including data from different cohorts during model training, while preserving patient privacy and adhering to data governance requirements through Federated Learning. In this work, however, we employed Macenko's color stain normalization algorithm. This approach was later refined through the use of neural networks in Kajetan Husiatyński's Master's thesis, *Neural Style Transfer Methods for Histopathological Image Analysis*, carried out in our lab as part of this project [1].

## 7.2 Chapters 3 and 4: Predicting therapy response from biopsy slides in AGEJ patients and CSCC patients

The technical similarities and biological differences between these two projects present an opportunity to discuss them together. AGEJ is an aggressive disease, with symptoms typically appearing at advanced stages, making early detection difficult. Moreover, available treatments, whether based on radiotherapy or chemotherapy, have limited success rates. In contrast, CSCC tumors can often be surgically removed, and in most cases, this treatment is definitive. However, it remains unclear why a minority of patients later experience recurrence or metastasis. While the prediction performance in the AGEJ study reached an AUROC of 0.80, CSCC progression could be predicted with an AUROC of up to 0.92 (in the local training scenario). There appears to be a positive correlation between the ease of treatment and classification performance: outcomes in the more aggressive disease are harder to predict. It goes without saying that, although different neural models were employed in these projects, every design choice was guided by the goal of achieving the highest possible level of reliability and effectiveness.

A key objective shared by these projects was not merely training a black-box classification algorithm but also gaining insights into the underlying disease mechanisms that may influence treatment response. Consequently, considerable emphasis was placed on model explainability. This was relatively straightforward in the HER2 study, where the models attention mechanism assigned a score to each image patch, making the classifier inherently interpretable by design. In contrast, the AGEJ and CSCC projects employed transformer classifiers, where the number of attention scores scales quadratically with the number of image patches and is further multiplied by the number of layers and attention heads. Moreover, attention scores are not the sole factors driving the classification decisions of transformers. These models require ad hoc explainability techniques to shed

light on their internal decision-making processes. We therefore opted for the Integrated Gradients algorithm, which can be applied to arbitrary neural architectures and assigns a single scalar value to each input instance, producing one heatmap per slide as desired.

Algorithmic explanation has a fractal nature: we strive to interpret a model, by means of an explainability algorithm, whose results need to be interpreted themselves. At some point, human judgment needs to step in. We did this latter interpretation in two complementary ways. First, we quantified the high attribution patches on a cellular level, with the aid of nuclei instance segmentation models and feature engineering. Second, since this approach is limited to quantifying image patches in isolation, we visually analyzed the attribution heatmaps to identify potential patterns.

While a single handcrafted feature can carry a clear, intuitive meaning, the complete set of 500 computed features is not easily interpretable. A more refined curation and selection of illustrative features was therefore necessary. Given our large populations (of image patches), most of the features differed significantly between the two groups, including those that had nearly identical distributions. The Common Language Effect Size turned out to be a very handy tool to guide our analysis.

Interestingly, in both diseases, image patches associated with positive outcomes tend to exhibit tumor cell populations with higher spatial autocorrelation than those linked to negative outcomes. However, while highly eccentric tumor cells are associated with therapy response in AGEJ, they are linked to disease progression in CSCC. A counter-intuitive observation in the AGEJ case is that non-responder-associated patches contain fewer lymphocytes than those associated with therapy response. Closer examination revealed that this pattern is present throughout the dataset, not just in regions of high attribution.

The CSCC study posed the additional challenge of working with a dataset composed of multiple cohorts. In the HER2 case, this issue was addressed by training a local model and applying stain color normalization to the external data. In the CSCC project, however, we tackled this problem by training the classifier with Federated Learning. This approach was later studied by Jakub Zacharczuk in his Master's thesis, *Federated Learning for Decentralized Model Training in Skin Cancer Histopathology*, conducted in our laboratory [2].

## 7.3 Chapter 5: WSI classification with language models

This chapter is, admittedly, a somewhat convoluted combination of ideas and observations, as outlined in its introduction. Its origins, however, were inspired by the wave of open-source models, collaborative research, and modular design that flourished within the NLP community following the popularization of the transformer architecture, particularly the Bidirectional Encoder Representations from Transformers (BERT) [3]. The development of numerous open-source models based on BERT (including the vision transformer) provided NLP researchers and practitioners with an unprecedented level of

flexibility, enabling them to take, modify, fine-tune, and interchange models with ease. This openness accelerated the field's progress by shifting the focus from the tedious task of neural network construction to experimentation and application. The primary motivation behind this work was to replicate that level of experimental freedom with transformer models in the context of WSI classification. The improvement gained from text-based pre-training came (almost) for free: if we can readily test different transformer architectures, why not apply those already pre-trained on text? This decision was not arbitrary. While the idea of using language pre-training to enhance performance in non-text domains may seem unconventional, it has been and continues to be explored in deep learning research [4].

At the time of writing, there are expectations from part of the AI community to produce artificial general intelligent agents based on large language models — models trained on text datasets and transformer architectures orders of magnitude larger than those explored in this work [5]. The core hypothesis of these efforts is that large transformers, if pre-trained on language, could do any type of intelligent task beyond routine text processing, including tasks that involve reasoning and planning. There is a conceptual similarity between this hypothesis and the ideas explored in our work, yet the goals are fundamentally different. We do not seek to endow our models with broad, human-like intelligence; rather, we aim to leverage the structural patterns present in natural language data to pre-train more effective visual classifiers. We believe that the patterns learned while modeling language can transfer to other data modalities, including histopathology images. In this sense, the linguistic nature of the data is both essential and, paradoxically, incidental. One possible alternative could be, for example, pre-training the transformer classifier with synthetic data [6].

Enabling WSI processing with deep transformers posed significant computational challenges, primarily due to the poor scaling of memory and compute requirements with increasing input size. The growing prominence of large language models has sparked interest in developing techniques for more efficient training and fine-tuning, now commonly referred to as parameter-efficient fine-tuning. These methods make it possible to adapt large transformers, pre-trained on high-performance computing infrastructure, using computational resources more readily available to individual researchers, such as a single GPU. In our case, we exploited the observation that the relevant visual information within a tissue slide can be condensed into a representation far smaller than the full set of image patches. Additionally, we took advantage of the fact that re-training only the normalization layers is often sufficient to successfully alter a models behavior, even in randomly initialized neural networks [7].

The following and final chapter of this dissertation continues along the line of incorporating domain-specific knowledge of visual patterns to improve digital pathology algorithms.

## 7.4 Chapter 6: Masked Context Modelling and Knowledge Distillation

There are many fundamental concepts in vision, both human and machine, which consistently reappear in the design of algorithms, and they are valid for histopathology imagery as well. The high resolution of WSIs makes them naturally suited for pyramidal representations [8], a strategy already employed by the codecs of their image formats. Multiple instance learning leverages "bag-of-words" representations [9]. Visual signals tend to be redundant, and the previous chapter takes advantage of this fact which applies to tissue slides as well [10]. Here, we designed an algorithm that exploits the information in the context of individual image patches, based on the spatial continuity of tissue along large image regions. It is intriguing to continue the study of how visual information processing can be incorporated into digital histopathology algorithms.

In this work, we propose an additional feature extractor fine-tuning stage with a novel task. It is useful to draw a parallel with multi-stage training of large visual-language models [11]. Training of such models begins with general image-text prediction tasks such as transcribing text from images, and they grow in abstraction as training progresses, finalizing with a high-level, domain-specific task, such as visual question-answering from corporate documents. Although the data always consists of paired images and text, this sequential approach bridges the semantic gap between the initial pre-training stage, which learns broad image-text patterns from abundant web-scale datasets, and the final fine-tuning stage, which relies on smaller, curated datasets for a specific domain. We observe a similar phenomenon in the MIL framework: the feature extractor is initially pre-trained on a large dataset to produce general-purpose image representations, while the MIL algorithm is subsequently trained on a task-specific histopathology dataset. The proposed MCM task is intended to mitigate this domain gap.

An indication that this project is going in the right direction is the surprising result is that the fine-tuned ResNet18 outperforms the larger EfficientNetV2-L teacher in some MIL tasks. This suggests that context-aware fine-tuning does not just copy teacher embeddings but learns a compressed, task-specific representation. This is aligned with our goal of improving the representations of the feature extractor in the context of specific downstream MIL tasks. Knowledge distillation is a fascinating topic in DL and it is still not understood in depth. While it is commonly used as a method of model compression, our motivation here is to avoid doing the context modeling task in pixel space.

## 7.5 Conclusion and outlook

The body of research presented in this dissertation demonstrates the potential of deep learning methods for histopathology image analysis, more specifically showcasing applications of biomarker prediction and therapy response assessment, as well as algorithmic strategies and novel approaches to address the unique challenges posed by this

domain. In the application chapters, deep learning techniques are employed for a variety of purposes: automating histopathologists' routine work, gaining insights into the inner mechanisms of diseases, or overcoming patient privacy and data governance drawbacks. Our work underlines the importance of domain-specific considerations, from data preparation to model explainability, and highlights how insights derived from these models can potentially support clinical decision-making.

In medical practice, it is a must to integrate data from different sources and modalities in order to make an accurate patient profile and tailor their treatment. This thesis, being about histopathology images, puts a strong emphasis on the processing of their visual signals. A promising direction that emerges from the presented works and their discussion is forging a closer dialogue in this interdisciplinary research. Just like the application of machine vision can aid in cancer research, posing meaningful, biological questions that can be answered by visual information processing can make these computational methods thrive. If we root our algorithms in clinically significant questions, that are simultaneously grounded in the application of these algorithms, we can open new pathways toward more accurate diagnoses, personalized treatments, and a deeper understanding of cancer.

## References

[1] Kajetan Husiatyski. "Neural Style Transfer Methods for Histopathological Image Analysis". Master's thesis. Warsaw, Poland: University of Warsaw, 2022.

[2] Jakub Zacharczuk. "Federated Learning for Decentralized Model Training in Skin Cancer Histopathology". Master's thesis. Warsaw, Poland: University of Warsaw, 2024.

[3] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[4] Wenlong Huang et al. "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents". In: *International conference on machine learning*. PMLR. 2022, pp. 9118–9147.

[5] Sébastien Bubeck et al. "Sparks of artificial general intelligence: Early experiments with gpt-4". In: *arXiv preprint arXiv:2303.12712* (2023).

[6] Abdul Fatir Ansari et al. "Chronos: Learning the language of time series". In: *arXiv preprint arXiv:2403.07815* (2024).

[7] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. "Training BatchNorm and Only BatchNorm: On the Expressive Power of Random Features in CNNs". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[8] Peter J Burt and Edward H Adelson. "The Laplacian pyramid as a compact image code". In: *Readings in computer vision*. Elsevier, 1987, pp. 671–679.

[9] Sivic and Zisserman. "Video Google: A text retrieval approach to object matching in videos". In: *Proceedings ninth IEEE international conference on computer vision*. IEEE. 2003, pp. 1470–1477.

[10] Fred Attneave. “Some informational aspects of visual perception.” In: *Psychological review* 61.3 (1954), p. 183.

[11] Lucas Beyer et al. “PaliGemma: A versatile 3B VLM for transfer”. In: *arXiv preprint arXiv:2407.07726* (2024).

Appendix A

# Supplementary Material for Chapter 2

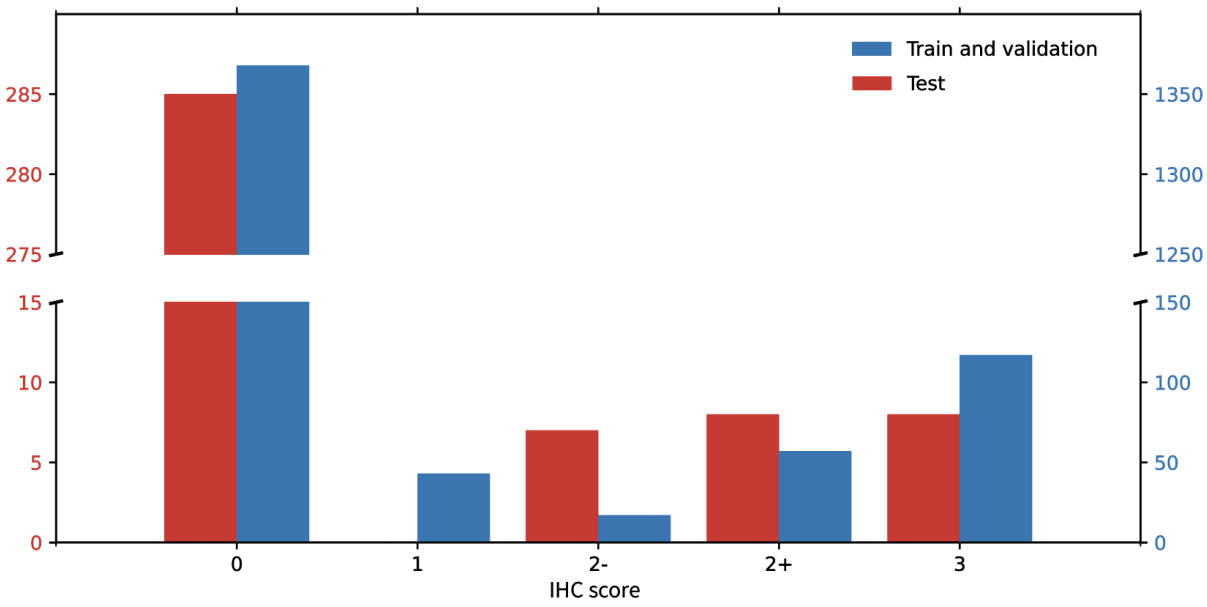
# Supplemental Material

## Predicting the HER2 status in oesophageal cancer from tissue microarrays using convolutional neural networks

Juan I. Pisula, Rabi R. Datta, Leandra Börner Valdez, Jan-Robert Avemarg, Jin-On Jung, Patrick Plum, Heike Löser, Philipp Lohneis, Monique Meuschke, Daniel Pinto dos Santos, Florian Gebauer, Alexander Quaas, Christiane J. Bruns, Axel Walch, Kai Lawonn, Felix C. Popp and Katarzyna Bozek

**Supplemental Table 1**. Staining patterns used by pathologists to assess the IHC score of HER2 stainings in biopsies. This analysis method was used because TMAs resemble biopsies more than whole slides.

| Score | Pattern of IHC staining for HER2 | HER2 status |
|---|---|---|
| **0** | No reactivity or membranous reactivity in any (or <5) tumor cell(s) | ***negative*** |
| **1** | Tumor cell cluster with a very weak membranous reactivity (at least 5 tumor cells) | ***negative*** |
| **2** | Tumor cell cluster with a weak to moderate complete, basolateral or lateral only membranous reactivity (at least 5 tumor cells) | ***equivocal*** **(ISH assessment required)** |
| **3** | Tumor cell cluster with a strong complete, basolateral or lateral only membranous reactivity (at least 5 tumor cells) | ***positive*** |



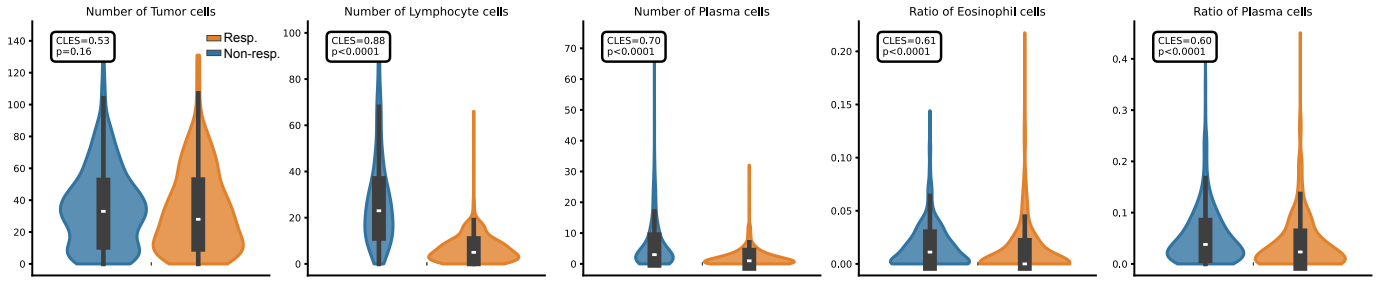

**Supplemental Figure 1**. IHC score distribution of our in-house datasets (with score 2 separated by positive and negative HER2 status).

Appendix B

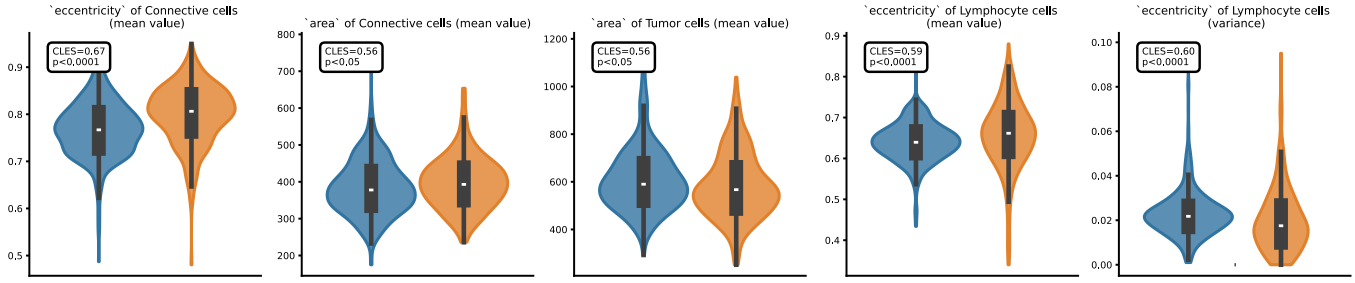
# Supplementary Material for Chapter 3

# Supplementary material

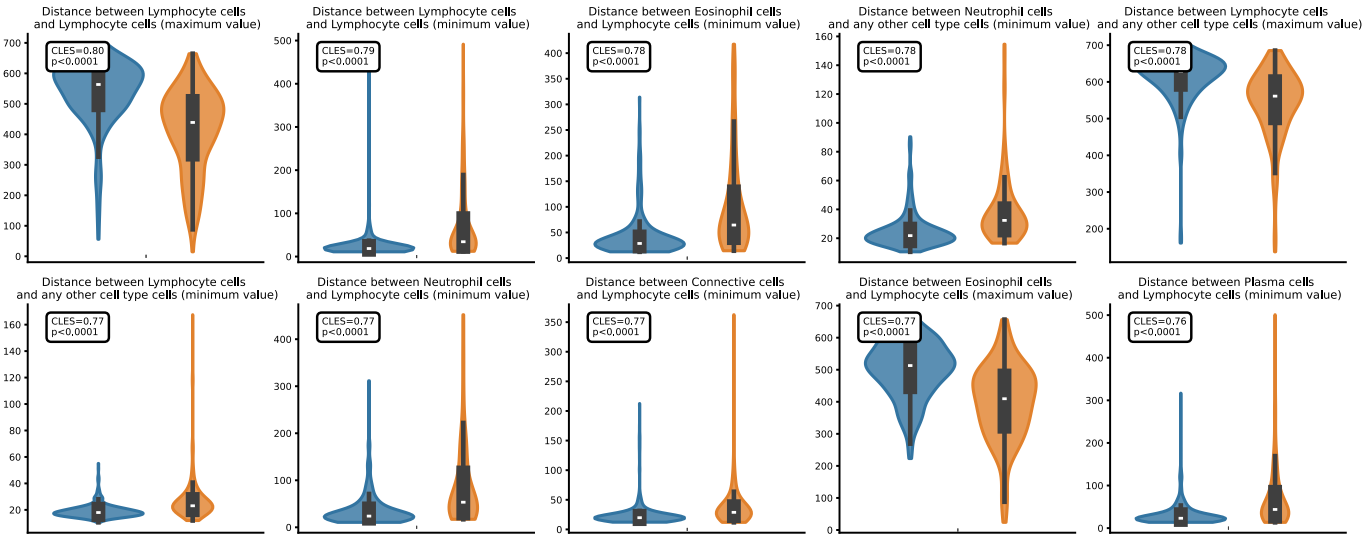
## Cell populations



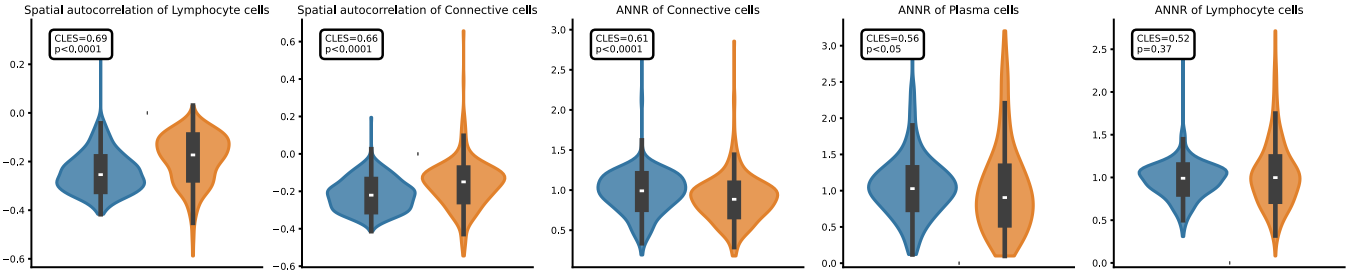

## Nuclei morphology



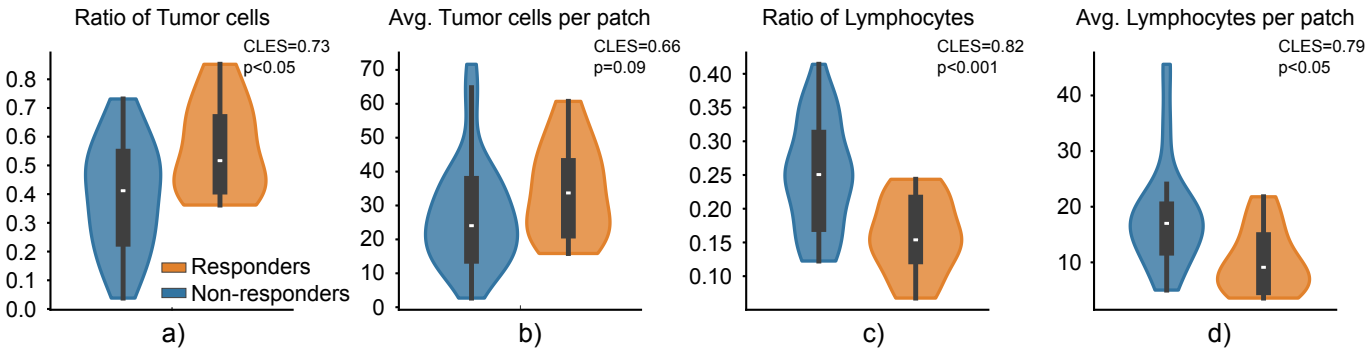

## Nuclei distances




## Spatial distribution




**Suppl. Fig. 1.** Additional selected features, split by their category.




**Suppl. Fig. 2.** Tumor cell and lymphocytes counts computed on complete tumor regions instead of highly attributed patches.

Appendix C

# Supplementary Material for Chapter 4

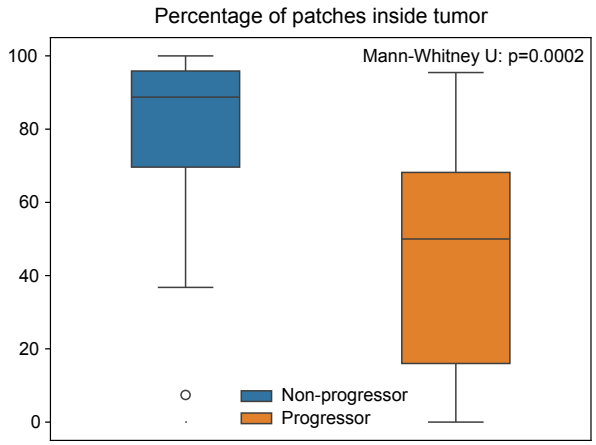
## Supplementary material

**A**

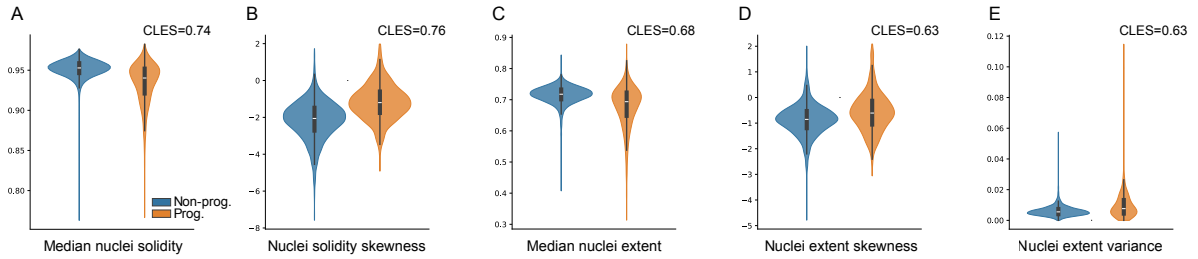
| Variable | | N | Odds ratio | OR (95% CI, univ.) | p-value (univ.) |
|---|---|---|---|---|---|
| **Vascular invasion** | No | 153 | | | |
| | Yes | 4 | | **9.41 (1.17, 193.30)** | **0.055** |
| **Perineural invasion** | No | 148 | | | |
| | Yes | 9 | | **6.71 (1.68, 33.12)** | **0.009** |
| **Invasion beyond subcutaneous tissue** | No | 148 | | | |
| | Yes | 9 | | **4.04 (1.02, 17.09)** | **0.046** |
| Diameter (cm, log10) | | 106 | | 2.30 (0.62, 9.39) | 0.227 |
| **Thickness >6mm** | No | 111 | | | |
| | Yes | 26 | | **2.27 (0.86, 5.73)** | **0.087** |
| Ulcer | No | 131 | | | |
| | Yes | 26 | | 2.10 (0.84, 5.08) | 0.101 |
| **Grading > 1** | No | 95 | | | |
| | Yes | 53 | | **1.97 (0.94, 4.15)** | **0.073** |
| Margin (cm, log10) | | 81 | | 1.64 (0.27, 10.39) | 0.590 |
| Infundibulocystic features | No | 154 | | | |
| | Yes | 3 | | 1.47 (0.07, 15.80) | 0.754 |
| Cornu cutaneum | No | 154 | | | |
| | Yes | 3 | | 1.47 (0.07, 15.80) | 0.754 |
| Keratoakanthoma | No | 134 | | | |
| | Yes | 23 | | 0.57 (0.16, 1.65) | 0.340 |



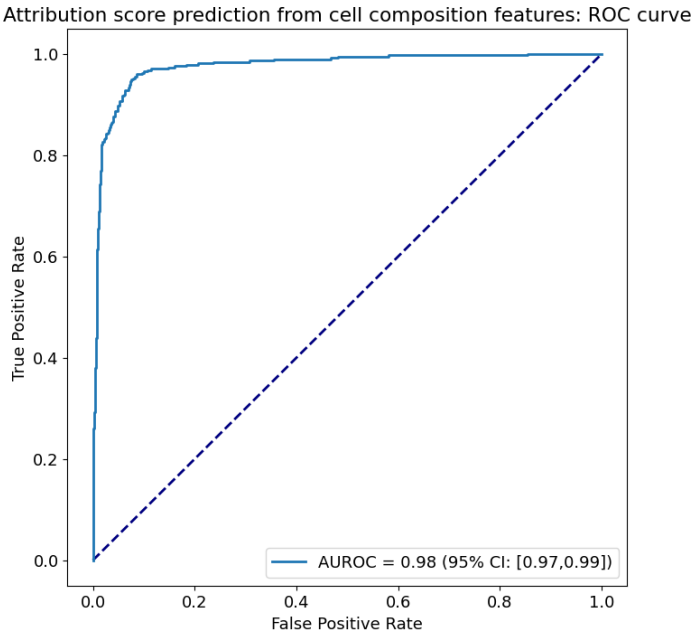

**Suppl. Figure 1:** Association of clinico-pathological parameters with cSCC progression risk calculated using logistic regression for Cologne patients with available data. Shown are Odds ratios (ORs) with 95% Confidence intervals (CIs) and univariate p-values.



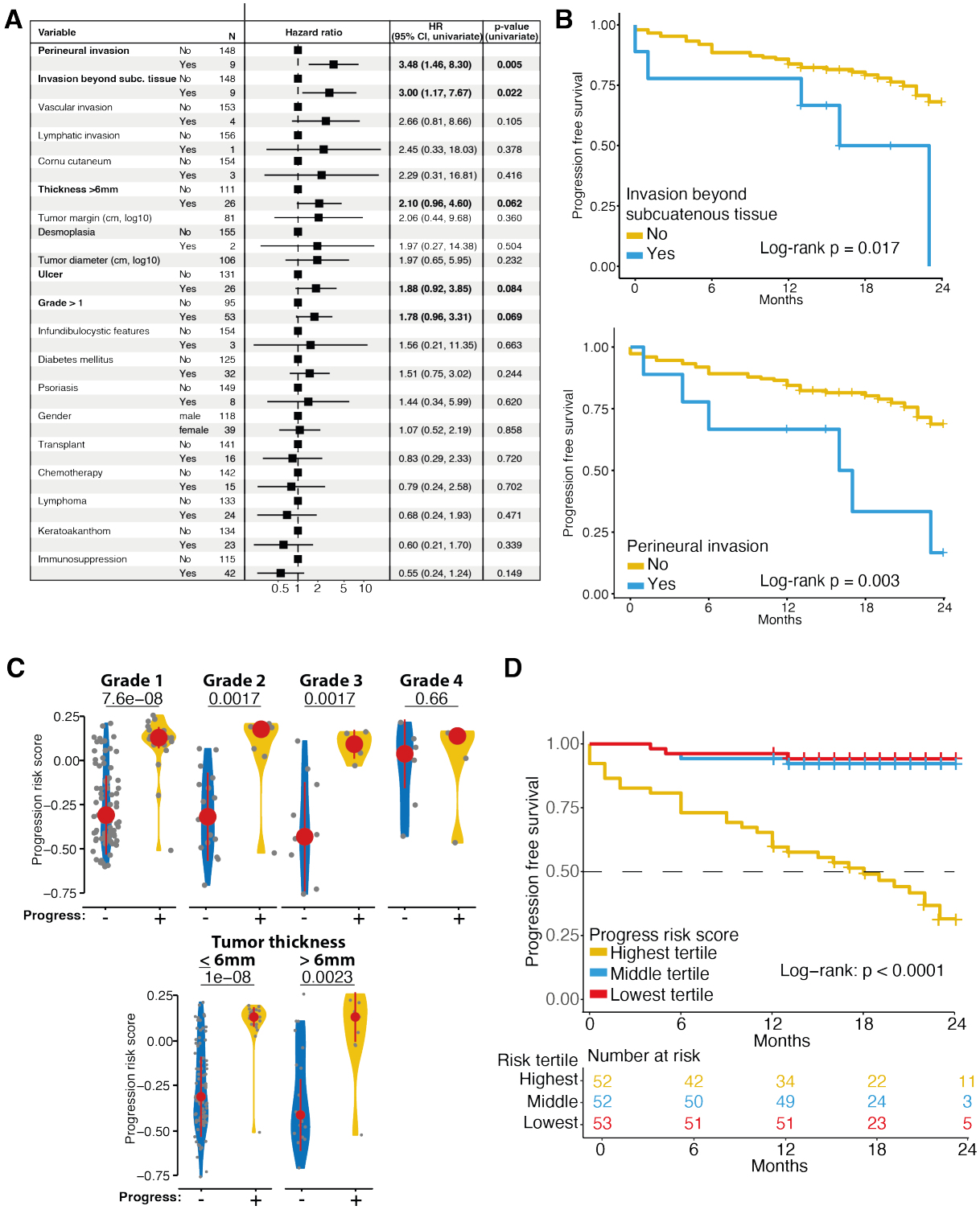

**Suppl. Figure 2:** Percentage of relevant patches (as detected by IGs) of individual patients inside the tumor regions. On average, non-progressors have more relevant patches inside the tumor compared to progressors.




**Suppl. Figure 3:** Violin plots of 5 tumor cell nuclei morphological features. Non-progressors have larger values of morphological solidity and extent (larger median, negatively-skewed distributions, **A-D**), while morphological extent has a larger variance in tumor cells from progressors **(E)**. All features are significantly different in both groups, with p-values < 0.0001 using Mann-Whitney U test.




**Suppl. Figure 4:** ROC curve of the XGBoost patch-level progression status classifier using cell-based features as input.




**Suppl. Figure 5: A:** Univariate association of clinico-pathological parameters derived from medical records & pathology reports with progression free survival of Cologne cSCC patients. Shown are Hazard ratio (HR) with 95% confidence interval (CI) based on Cox proportional hazard models. N indicates number of patients with available data per category. **B:** Kaplan-Meier curves for Cologne patients with or without invasion beyond subcutaneous tissue (top) or perineural invasion (bottom). **C:** Comparison

of deep learning-based progression risk scores in Cologne patients with or without cSCC progression stratified by pathological grade (top) or tumor thickness >6mm (bottom). Shown are median and median absolute deviation. p-values calculated by t test. **D:** Progression free survival of patients grouped into tertiles of the deep learning-based progression risk score.

**Suppl. Table 1**

Top 100 features with the largest CLES, or probability of superiority, between the groups. To avoid displaying redundant features, pairs of features with a Pearson correlation coefficient bigger than 0.9 are grouped together, and a single feature from the group is shown. The rows are sorted in descending order of CLES for each feature type. The “Higher in” column indicates the group with larger feature values. The fraction of image patches that do not show any value for the features are shown, and features missing in more than 90% of the patches are not displayed. All the features in the table are significantly different in both groups with $p<0.0001$ using Mann-Whitney U test. Description of nuclei morphology features can be found in the documentation of `skimage.measure` (https://scikit-image.org/docs/stable/api/skimage.measure.html#skimage.measure.regionprops).

Appendix D

# Supplementary Material for Chapter 5

# Supplementary material

**Supplementary Table 1**: Performance of different MIL algorithms in LNM and IBC classification tasks using CTransPath features [6]. Best and second best classification results are in **bold** and underlined, respectively.

| Method | x10 magnification | | x20 magnification | |
|---|---|---|---|---|
| | 25% train set | 100% train set | 25% train set | 100% train set |
| **Lymph Node Metastases classification** | | | | |
| ABMIL [1] | **0.881** | 0.910 | **0.938** | 0.971 |
| CLAM [3] | 0.695 | **0.928** | 0.717 | 0.953 |
| DS-MIL [2] | 0.541 | 0.744 | 0.521 | 0.934 |
| TransMIL [4] | 0.634 | 0.870 | 0.654 | 0.935 |
| Wagner et al. [5] | 0.663 | **0.928** | 0.757 | **0.974** |
| **Ours** | 0.512 | 0.783 | 0.448 | 0.957 |
| **Invasive Breast Carcinoma subtype classification** | | | | |
| ABMIL [1] | 0.868 ± 0.062 | 0.896 ± 0.066 | 0.690 ± 0.298 | 0.893 ± 0.057 |
| CLAM [3] | **0.921 ± 0.065** | 0.929 ± 0.033 | 0.895 ± 0.051 | **0.937 ± 0.300** |
| DS-MIL [2] | 0.913 ± 0.056 | 0.934 ± 0.037 | **0.903 ± 0.050** | 0.934 ± 0.036 |
| TransMIL [4] | 0.890 ± 0.060 | 0.934 ± 0.043 | 0.882 ± 0.061 | 0.924 ± 0.042 |
| Wagner et al. [5] | 0.903 ± 0.059 | **0.935 ± 0.035** | 0.881 ± 0.061 | 0.927 ± 0.052 |
| **Ours** | 0.860 ± 0.087 | 0.928 ± 0.043 | 0.860± 0.062 | 0.914 ± 0.057 |

# References


[1] Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: Dy JG, Krause A (eds) ICML 2018, PMLR, vol 80. PMLR, pp 2132–2141

[2] Li B, Li Y, Eliceiri KW (2021) Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR 2021. Computer Vision Foundation / IEEE, pp 14318–14328

[3] Lu MY, Williamson DF, Chen TY, et al (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering 5(6):555–570

[4] Shao Z, Bian H, Chen Y, et al (2021) Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Ranzato M, Beygelzimer A, Dauphin YN, et al (eds) NeurIPS 2021, pp 2136–2147

[5] Wagner SJ, Reisenbüchler D, West NP, et al (2023) Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell 41(9):1650–1661.e4. https://doi.org/https://doi.org/10.1016/j.ccell.2023.08.002, URL https://www.sciencedirect.com/science/article/pii/S1535610823002787

[6] Wang X, Yang S, Zhang J, et al (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis 81:102559. https://doi.org/https://doi.org/10.1016/j.media.2022.102559, URL https://www.sciencedirect.com/science/article/pii/S1361841522002043