Proposal: Machine Learning for Histological Analysis in Pathology

Caden Steele CADENS090@GMAIL.COM
Gianforte School of Computing
Montana State University
Bozeman, Montana

Editor: None

Abstract

Advancements in medical imaging are revolutionizing the field of oncology, yet the analysis of histological slides remains a bottleneck due to the labor-intensive and highly specialized expertise required. Whole Slide Images (WSIs), comprising high-resolution data, present immense opportunities for cancer research and diagnosis but pose significant challenges in manual interpretation and analysis. This proposal outlines the development of a cutting-edge machine learning tool, leveraging the innovative Cerberus model architecture, to automate tissue segmentation and classification tasks for WSIs accessed through the Genomic Data Commons (GDC) Analysis Center. The Cerberus model employs a modular, fully convolutional network (FCN) architecture, featuring a shared encoder and task-specific decoders, to achieve scalability, generalization, and multi-task learning. By integrating transfer learning, the model facilitates the seamless addition of new tasks without the need for full retraining, reducing computational overhead and enabling continuous adaptation to emerging datasets. This tool will initially focus on segmenting and classifying tissues from colorectal, breast, and oral cancer samples, but the design is optimized for additional tasks. The proposed tool offers accuracy and efficiency on par with expert pathologists under time constraints, promising to enhance diagnostic precision, accelerate research, and reduce the workload of medical professionals. By investing in this initiative, the GDC can establish itself as a global leader in automated histopathological analysis, advancing cancer diagnosis and research through innovative solutions.

1. Introduction

Medical imaging advances almost as quickly as the cancer rates in our American citizens. Despite this, medical research remains slow and expensive. But with diligent work, our technology marches forward and our collective knowledge expands. One of the many bottlenecks in research is the examination and analysis of histological slides. Requiring a seasoned pathologist, the work is time-consuming, and requires years of schooling. Streamlining the work in an accessible format would benefit researchers, educators, students, hospitals, and the private sector alike.

Whole slide images (WSI) are a series of images captured at different magnifications, and stitched together into a single, extremely high resolution image. A WSI being 100,000x100,000 pixels is not uncommon. WSIs are big, up to 2 or 3 gigabytes in size, and a project with thousands of WSIs necessitates months to analyze each image thoroughly. This tool could give a pathologist a starting point in a sea of WSIs, instead of sifting through each image blindly. It could allow a single pathologist to do the work of several at once. Allow research teams, without access to a pathologist, to continue their projects or launch new ones. Even allowing an overworked hospital pathologist to make a quality biopsy report in time for their next patient.

We propose building a tool for the Genomic Data Common's (GDC) Analysis Center which uses cutting-edge machine learning (ML) technology to analyze WSIs at a comparable level to humans. Our tool will use a pre-trained fully convolutional network (FTN) architecture to read a WSI file (.SVS), or a cohort of them, from the GDC Data Portal, and automatically perform two specific tasks: tissue segmentation and tissue classification.

Leveraging the newly created Cerberus model (Graham et al., 2023), we can build this tool with generalization and future scalability in mind. Despite launching with only two tasks and a handful of tissues, this model allows us to continually add new tissue and cancer types without the long, processing-intensive issue of re-training. With maintenance and upkeep, this initial architecture could let the GDC potentially host the world's leading WSI analysis technology, and one of the first organizations to possess software capable of analyzing a WSI of every tissue and cancer existing within the human body without a colossal processing overhead.

1.1 System

A new system architecture, dubbed the Cerberus model and created in Graham et al. (2023), uses FCNs to create a generalizable architecture capable of learning to perform a myriad of separate tasks. It involves using two FCN's, one as encoder and at least one more as a decoder. This is a similar structure to a transformer, but lacks most of the skip connections. The innovation this model establishes, is that a system of one primary encoder with several decoders, each assigned to a particular task, is viable. Multi-task learning (MTL) has been difficult thus far, but Cerberus has proven to be reliable.

Traditionally, one neural network (NN) is trained to be very skilled at performing one specific task, such as predicting the number of fruit a tree will produce or classifying a species cat based on paw size. The Cerberus model breaks that mold by being generalizable to several tasks. Instead of building an ensemble of networks for different tasks, we can build one primary encoder and then build one decoder for each task. By employing multi-task learning, Cerberus massively reduces the complexity, training time, and build time needed. Our architecture will use two decoders: one for tissue segmentation and one for cell classification.

Image segmentation is a general term for discerning objects within an image. Class segmentation is discerning objects of different types, while instance segmentation is discerning multiple of the same object. Our software will take a whole slide image (WSI), or a cohort of several, stored in .SVS files from the GDC's database, and segment cancerous tissue from normal tissue.

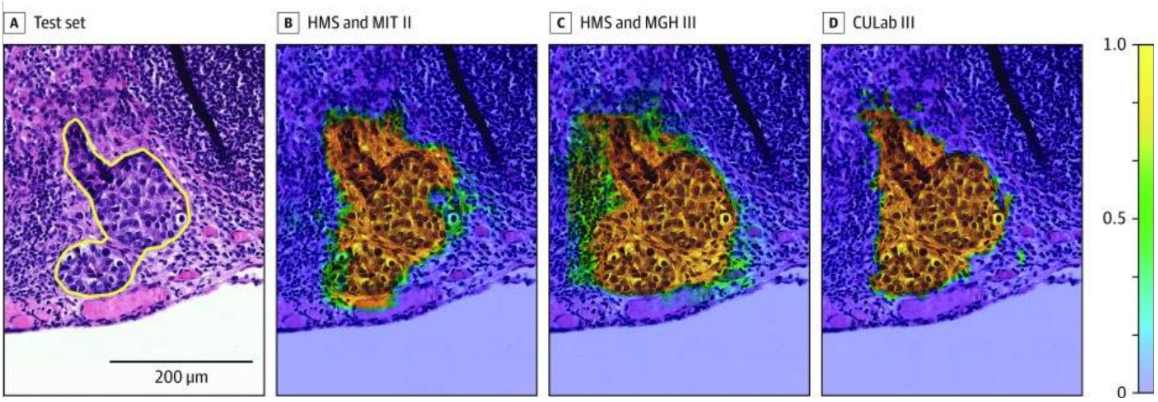

Figure 1: Image segmentation of cancer metastases in lymph nodes (B et al., 2017)

Classification is a common use for FCN's. By taking a detailed look at an image, it will assign a label to the image or a subsection within it. We plan to use this to analyze the types of cells present in a WSI and determine the tissue of origin. This could be extended in the future to determine the tissue source, such as an organ or region of the body.

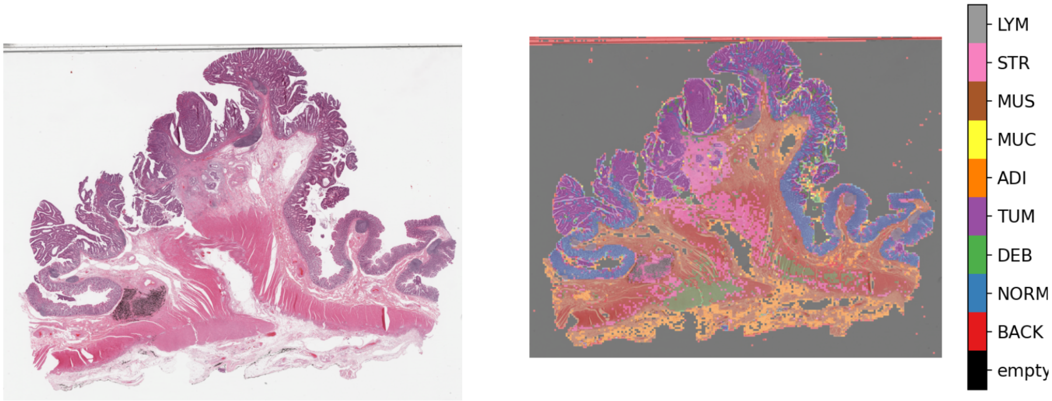

Figure 2: Before and after of tissue classification by cell type. (PyTorch)

A second benefit of Cerberus is cross dataset training. Datasets annotated for a particular purpose (i.e. a set of cell images annotated to denote cell type, but nothing else) are generally unusable for other tasks. However, since Cerberus uses the same encoder for multiple tasks, features from every dataset are extracted into the encoder. This means Cerberus can use features unrelated to a current task to enhance its performance. Therefore, generalizing the system can actually increase performance, whereas in most other systems this would not be true.

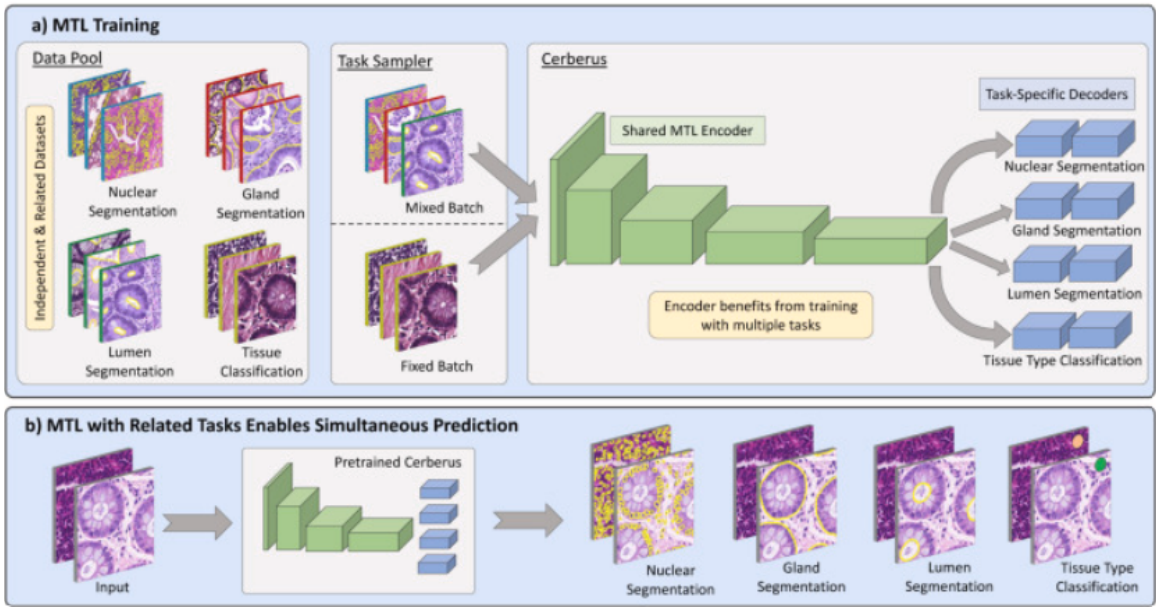

Figure 3: Cerberus Model (Graham et al., 2023)

This architecture is highly conducive to transfer learning, especially when given a task similar to previous tasks. In Graham et al. (2023), they use transfer learning to efficiently create a subtyping decoder. Their model classifies cell types, and using transfer learning, train a decoder which then classifies nucleus types. Given one image as input, the architecture can use both decoders to classify cell and nucleus types. This new nucleus decoder can then be used in transfer learning to create a decoder which segments a nucleus from the surrounding cell. Thus, we can create a chain reaction of scalability and utilization. This is highly scalable, but it is not known if an upper limit on this kind of generalizing exists or what drawbacks could eventually emerge.

We plan to have segmentation and classification functionality available for the following body regions and some associated tissues.

Tissues to be classified:

1. **Colon**: adipose, lymphocytes, mucus, smooth muscle, normal colon mucosa, inflammatory, cancer-associated stroma, colorectal adenocarcinoma epithelium

2. **Oral Cavity**: epithelium, oral squamous cell carcinoma

Tissue to be segmented:

1. **Breast**: tumor, stroma, lymphocytic infiltrate, glandular secretions, blood, exclude, metaplasia NOS, adipose, plasma cells, other immune infiltrate, mucoid material, normal acinus or duct, lymphatics, nerve, skin adnexa, blood vessel, angioinvasion

These tissues and cancers are varied, and are annotated for different tasks. The design of Cerberus will allow us to use this differentiation to our advantage and increase the robustness of the predictions. In the future, these datasets could be used for additional tasks, such as classifying the stage of a cancer or assigning a Gleason grade to prostate cancer.

2. Literature Review

Machine learning (ML), particularly deep learning, has increasingly gained prominence in the field of digital pathology for its ability to analyze and interpret complex histological images. Whole slide images (WSI), generated from high-resolution scanning of histology slides, provide rich data crucial research, education, and diagnoses. However, the storage space consumption and complexity of WSI's challenge manual interpretation, paving the way for machine learning driven analyses to assist pathologists in achieving accurate and consistent results. This literature review examines the applications and feasibility of ML techniques in analyzing WSI's and pathology.

2.1 Viability

According to B et al. (2017), comparing segmentation machine learning algorithms to a panel of accredited pathologists proved FCN and CNN algorihtms to be viable. They compared a group of pathologists with a time constraint (WTC) and a group without time constraints (WOTC), to the algorithms on identifying metastases in lymph nodes. Against the pathologists WOTC, the algorithms performed about 10% worse. While against the pathologists WTC, the algorithms performed similarly or better. Further breaking down results, Graham et al. (2023) notes that when identifying micrometastases specifically, the best performing pathologist missed 37% of instances, but the algorithms typically did not.

Recent advancements in convolutional neural networks (CNN) have enabled significant progress in the analysis of WSI's. CNN's, with their ability to capture spatial hierarchies and recognize intricate patterns, have been pivotal in streamlining pathology for tasks such as cell segmentation, tumor classification, and biomarker detection. For example, A et al. (2017) demonstrated the effectiveness of CNN's for skin cancer classification, achieving performance comparable to that of dermatologists in dermatopathology. This study demonstrates the potential for similar applications in cytopathology, histopathology, neuropathology and more.

2.2 Technical Challenges

Despite the significant benefits to using machine learning in classification and segmentation, some challenges remain. Differences in staining practices will be the most likely culprit of false positives. Changes in tissue staining for histological analysis may occur due to differences in purpose, region, institute, or decade. Large differences like these ceate a weakness in datasets and may bias any learning model. Several proposed solutions, such as stain normalization, have been tested and found to be lacking (Hoque et al., 2024). We propose grayscaling in our preprocessing to combat this problem, but there remains no perfect solution.

Real life seldom mimics controlled training, and co-occuring pathogens are a variable we cannot account for. At this time, there are little to no datasets containing images of co-occuring pathogens, such as a tumor with an infection. The technology is perfectly capable of accounting for these complications, however, we would need data on every type of cancer with every type of co-occuring pathogen to accurately factor these pathogens into the model. Data is the limiting resource in this problem. We recommend investing into data collection on this topic in the future.

3. Methodology

The methodology describes training the model, so it can be used as a pretrained model by end users. The techniques described will give us a robust and scalable architecture with a high degree of accuracy. When more datasets are added to detect more cancer types or mutations, this will become an iterative process. Due to the architecture's modularity, new techniques can be added and obsolete techniques removed with ease. The modularity also allows for systemic testing of hyperparamters for determining optimal settings.

3.1 Data Preprocessing

Before any data is processed in training, it must be preprocessed. By making adjustments to our data, we can optimize the data for processing. Many preprocessing techniques change the data such that it's unrecognizable for humans, but for computers, these changes can mean huge improvements in performance. The following preprocessing techniques used in Kumar et al., 2024 to help extract features and reduce complexity in a diverse dataset, had promising results.

1. **Grayscale conversion** is the process of converting an image's colorspace from RGB, CMYK, etc, to black and white.

2. **Otsu's Method**, also known as Otsu's Binarization, is an algorithm which classifies a pixel as foreground or background by minimizing intra-class intensity variance.

3. **Gaussian Denoising** is the process of removing noise using a Gaussian curve.

4. **Distance transform** is used to calculate the euclidean distance of each pixel to the nearest edge within the image for identification of important cells.

5. **Watershed Transformation** is a technique which mimics the flow of water to help segment different tissue structures.

3.2 System Architecture

Once the dataset has been preprocessed, it's ready to be for training the model. We will be using the Cerberus architecture for its ability to perform generalization and scalability.

The Cerberus model is created and detailed in (Graham et al., 2023), and most of the information in this section comes from here. The Cerberus model is a fully convolutional neural network architecture which uses one primary neural network as an encoder and at least one more as a decoder. The model can perform more tasks by simply adding more decoders. The power to give one network multiple tasks to perform is extremely significant. The lack of this ability is a major hindrance in most machine learning architectures.

Our encoder structure will use the DenseNet201 model. According to (Kumar et al., 2024), DenseNet201 and DenseNet121 are the best options for image classification when compared to other models. As the name suggests, it is dense and parameter heavy. However, its complexity benefits MTL and allows more feature extraction and therefore more scalability. Therefore, it emerges as the optimal choice for our Cerberus encoder. The decoder will use a U-Net model to allow upsampling until the dimensions of the input image are reached. By using a fully convolutional neural network (FCN) as the encoder and decoder, the architecture can use images of any size, eliminating a large blocker in current research.

3.3 Training

Once data has be preprocessed to make it easier to handle, the architecture is ready for initial training. The following steps are similar for all machine learning, but Graham et al. (2023), who created the Cerberus model, iterated through the steps about 90,000 times. As this architecture is designed for expanding its capabilities, further training needs to be done every time a task is added. Initially, we will train one encoder and two decoders. Every decoder, corresponding to a new task, will involve training the model again. Retraining the entire system for every task would be exponentially expensive in time and space, so transfer learning is used to remove the redundant learning for every subsequent task added.

1. **Initialize** with random values.
2. **Select task** randomly to feed data through.
3. **Sample data** with *mixed batch* sampling (Graham et al., 2023). A mixed batch contains data from multiple tasks.
4. **Feedforward data**, data information is fed through the architecture, where a result is created.
5. **Loss aggregation**, using a cross-entropy-based loss function for both segmentation and classification tasks, the model's results are used to score its performance.
6. **Backpropogation** is used to update weights based on the loss aggregation results.
7. **Dynamic freezing** locks sections of the architecture in place so they can't be updated by backpropogation. The encoder will never be frozen, but decoders will be frozen if no data related to its task is fed through.
8. **Transfer learning** is the process of using a pre-existing model as the base for another. In this case, when a new task is added, an old decoder is copied and used as the base for a new decoder.

3.4 Post Processing

While preprocessing changes the input data for training, post processing affects the already trained model. A fully trained or semi-trained model uses post processing techniques to hone accuracy without creating large changes.

1. **Hard Example Mining** is a process of taking false positive or negative instances and adding them to the learning set.

2. **Fine-Tuning**, more specific than dynamic freezing, is freezing a section of an inner model while training to isolate learning to a specific area. This is often used in transfer learning when the new task is particularly similar.

3.5 Evaluation

A test set, about 10%, is created by randomly picking a number of data samples from each task. This is fed through the architecture to evaluate the model's accuracy. The test set comprises diverse samples that represent varying histological patterns and image qualities. We used much of the same evaluation criteria as Graham et al. (2023).

Segmentation performance is measured with the Dice score and Panoptic Quality (PQ). The Dice score measures how well the foreground pixels are separated from the background pixels, but does not measure how well neighboring objects are segmented. The PQ measures the performance of instance segmentation. PQ is calculated for each image and averaged for total performance. For multi-class segmentation, when an object has multiple correct labels, PQ is calculated for each class and averaged.

Classification performance is measured with mean average precision (mAP) and mean F1 (mF1). AP and F1 are calculated by image, and an average over each class is calculated, giving mAP and mF1.

These measurements shall be given to an end user upon using this tool so one may decide for themselves how to use the data we provide.

4. Datasets

The following are the datasets planned to use in the training of this architecture.

1. **Kather100k**: 100,000 non-overlapping image patches of human colorectal cancer and normal tissue. The tissue types present are Adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, colorectal adenocarcinoma epithelium.
   Source: Kather et al. (2019)
   Data URL: https://zenodo.org/records/1214456

2. **CRC-TP**: 280k patches form 20 patients with colorectal cancer. The tissue types present are tumor, stroma, complex stroma, smooth muscle, benign, inflammatory, and debris.
   Source: Javed et al. (2020)
   Data URL: https://www.kaggle.com/datasets/haashaatif/crc-tissue-phenotyping-crc-tp-dataset

3. **Breast Cancer Semantic Segmentation**: Roughly 36,000 segmentation annotations of breast cancer tissue regions. Tissue types present are outside-roi, tumor, stroma, lymphocytic infiltrate, necrosis or debris, glandular secretions, blood, exclude, metaplasia NOS, fat, plasma cells, other immune infiltrate, mucoid material, normal acinus or duct, lymphatics, undetermined, nerve, skin adnexa, blood vessel, angioinvasion, dcis, other.
   Source:
   Data URL: https://www.kaggle.com/datasets/whats2000/breast-cancer-semantic-segmentation-bcss/data

4. **Histopathologic Oral Cancer Detection**: 1224 histological images of oral cavity normal epithelium and oral squamous cell carcinoma. The set is further split into 528 images at 100x magnification and 696 images at 400x magnification.
   Source: Rahman et al. (2020)
   Data URL: https://www.kaggle.com/datasets/ashenafifasilkebede/dataset/data

5. Budget

The *Details* column describes items involved in the creation of this project, subdivided when necessary. The *speculative* column shows costs based on outsourcing this project or building this project as a stand-alone product. The *Planned* column shows the likely cost of internally building this project as described in this document. The *Difference* column describes the monetary difference between the two previous columns.

Details	Speculative	Planned	Difference
Personnel			
Network Engineer (47.49 hr)	98,769.00	0	98,769.00
Software Engineer (77.67 hr)	161,552.00	46,602.00	114,950.00
UI Developer (42.00 hr)	87,362.00	25,200.00	62,162.00
Sub-Total	**347,683.00**	**71,802.00**	**275,881.00**
Data Acquisition			
Public Datasets	0 - 5,000	0	0
Commercial Datasets	10,000 - 1,000,000	0	10,000
Commissioned Datasets	50,000 - 5,000,000	0	50,000
Sub-Total	**60,000**	**0**	**60,000**
AWS Servers	141,649.20	0	141,649.20
Total Projected	**549,332.20**	**71,802.00**	**477,530.20**

5.1 Budget Justification

The numbers for personnel costs are based on the average annual salary for that position in the United States. The planned personnel costs use the position's average salary based on a 12 hour work week, as 12 hours is the expected weekly workload for this project. It should also be noted that the planned personnel costs of this project are measured in **opportunity cost** and not actual paid salary.

This project uses free open-source datasets, but large-scale usage may require paid or commissioned datasets. The costs are extremely variable depending on source, quality, recency, collection method, and ratio of the number of samples to number of individuals sampled, among other things.

The costs of renting an AWS server is substantially less than buying, building, and maintaining a dedicated server, so the likely annual price of renting a dedicated enterprise level AWS was used.

6. Conclusion

New advances in machine learning or medical image analysis can be measured by the day. This is an opportunity for the GDC to plant the seeds of a project with the infrastructure to become world class. The possibility of having a generalized AI that can classify every type of cell, tissue, cancer, or abnormality is near, and this is the foundation of such technology.

Using the Cerberus model, this architecture is built for scale and generalization. There is proof that the models can easily accommodate multiple types of cancers and tissues at one time. There is proof that the models can compare to human quality, or even surpass it under time constraints. Segmentation and classification software with the relatively small scale we plan for launch, is still invaluable to groups across the nation. An endorsement from the GDC would go far in legitimizing this technology, especially with the accuracy numbers shown in the literature. The number of hours spent by a pathologist pouring over images could be shaved down by a huge margin.

The competitive market cost of such a project being half a million dollars, as seen in the budget section, is quite realistic. Despite that, we are willing to complete this for an opportunity cost equivalent to about $70,000. This project should easily pass phase one based on usefulness, viability, scalability, and cost.

References

Esteva A, Kuprel B, Novoa RA, Ko J., Swetter SM, Blau HM, and Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. 546:115–118, 2017.

Ehteshami Bejnordi B, Veta M, Johannes van Diest P, and et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2155–2265, 2017.

Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, David Snead Fayyaz Minhas, and Nasir Rajpoot. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83, 2023.

Md. Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, and Tapio Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102, 2024.

Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis*, 63:101696, 2020.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *Plos Medicine*, 2019.

Y. Kumar, S. Shrivastav, and K. et al. Garg. Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification. *Sci Rep*, 14, 2024.

Tabassum Yesmin Rahman, Lipi B. Mahanta, Anup K. Das, and Jagannath D. Sarma. Histopathological imaging database for oral cancer analysis. *Data in Brief*, 29:105114, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2020.105114.