西安交通大學

硕士学位论文

透明细胞肾细胞癌病理图像分析及语义理解方法研究

学位申请人：Pargorn Puttapirat

指导教师：李辰 教授

学科名称：计算机科学与技术

2019 年 5 月

Semantic Understanding of Histopathological Images in Clear Cell Renal Cell Carcinoma

A thesis submitted to

Xi'an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Pargorn Puttapirat

Supervisor: Prof. Chen Li

(Computer Science and Technology)

May 2019

硕士学位论文答辩委员会

透明细胞肾细胞癌病理图像分析及语义理解方法研究

答辩人：Pargorn Puttapirat

答辩委员会委员：

西安交通大学教授：田锋

西安交通大学教授：刘均

西安交通大学教授：李辰

西安交通大学教授：张未展

西安交通大学副教授：王志文

答辩时间：2019 年 05 月 24 日

答辩地点：西安交通大学彭康楼 224

摘 要

随着癌症病例的逐渐增加，对准确的癌症筛查、诊断和治疗的需求也在不断增长，而癌症发病率只会随着世界人口的增长而不断上升。为了解决这个问题，数字病理学的发展利用了现代信息技术和先进算法的优势，如今已能够达到与现有人类医务人员相比近似或更好的水平。促进这种发展需要强大的生物医学数据基础设施以及大规模数据集和专业知识。通过标注称为全视野数字病理切片的扫描玻璃载片，数字组织病理学图像蕴含的丰富语义信息得以进行提取，而这需要能够处理这种类型生物医学数据的软件以及兼容现有病理学程序的数据处理流程。如今人们对大规模标注的组织病理学数据集的需求正在增加，因为人工智能技术的发展需要它们来推动自动诊断、大规模筛选和表型 - 基因型关联研究的发展。本论文提出了三个主要发现：一个开放的图像标注软件系统 OpenHI，用于高效地进行组织病理学图像的合作标注，并使标注数据能够包含标准化的语义和像素级的精度；使用提出的标注软件系统标注来自 TCGA KIRC 项目的透明细胞肾细胞癌（ccRCC）诊断切片；以及通过一个基于对现有分级指南进行解释而新提出的分析方法，对可行的组织病理学图像分析方法进行探索性分析。

该框架的即时响应处理算法可以执行大规模的组织病理学图像标注，并可作为数字病理学领域生物医学数据采集的基础。它可以进行进一步拓展以标注各种肿瘤类型的组织病理学图像。此框架已开源，用于标注 ccRCC 组织载玻片的 OpenHI 实现源码位于 https://gitlab.com/BioAI/OpenHI。文中讨论了用于模拟精确的虚拟放大倍数的最佳实践和方法。试验的标注结果表明，利用预定义的标注区域划分进行多专家标注能够在一定程度上符合两个标注者的判断标准。最后，一些选定的符合癌症分级指南描述的图像特征已在一个人工确定细胞核位置的自建数据集中得到测试。这些特征可以在基于 10 倍交叉验证的线性 SVM 分类器上达到 71.24％的细胞核级别分类准确度，以及在基于神经网络的分类器上达到 71.80％（SD = 2.06）的总体准确度。区分来自不同切片和不同患者的数据的实现也许能够达到更优的性能。

关 键 词：数字病理学；全视野数字病理切片；WSI；图像标注；数字切片；虚拟放大倍数；组织病理学；癌症诊断；癌症分级；基因型 - 表型关联；医学图像分析；透明细胞肾细胞癌；肾癌

论文类型：理论研究; 应用研究

ABSTRACT

A growing demand for accurate cancer screening, diagnosis, and treatment is the result of increasing number of cancer incidence which will only grow with growing world population. Tackling this problem, development in digital pathology taking advantages of modern era of information technology and advanced algorithms that may be able to perform at the level similar or better than existing human medical personnel. Facilitating such development needs a strong biomedical data infrastructure as well as large-scale dataset and expert knowledge. Consolidating semantically rich digital histopathological image by annotating scanned glass slides known as whole-slide images requires a software capable of handling this type of biomedical data and a support for procedures which align with existing pathological routine. Demand for large-scale annotated histopathological datasets are on the raise because they are needed for developments of artificial intelligence techniques to promote automatic diagnosis, mass screening, or phenotype-genotype association study. This thesis presents three main findings: an open annotation framework for efficient collaborative histopathological image annotation with standardized semantic enrichment at a pixel-level precision named OpenHI, a demonstration using the proposed annotation framework to annotate diagnosis slides of clear cell renal cell carcinoma (ccRCC) from TCGA KIRC Project, and exploratory analysis on feasible histopathological image analysis method with a newly proposed analysis approach based on the interpretation of existing grading guideline.

The framework's responsive processing algorithm can perform large-scale histopathological image annotation and serve as biomedical data infrastructure for digital pathology. It could be extended to annotate histopathological image of various oncological types. The framework is open-source and available at https://gitlab.com/BioAI/OpenHI. The implementation of OpenHI for the annotation of ccRCC tissue slides. Best practices and methods for simulating accurate virtual magnification has been discussed. Annotation results from the trial shows that utilizing pre-defined clusters for multi-expert annotation somewhat align the opinion of two annotators. Finally, selected image features that align with the description of grading guideline have been tested an inhouse dataset with manually localized nuclei. The features can achieve nuclei-level classification accuracy of 71.24% with 10-fold cross-validation on a linear SVM-based classifier and overall accuracy of 71.80% (SD = 2.06) on a neural network-based classifier. Further implementation at a slide-level and patient-level may yield better performance.

KEY WORDS: OpenHI; digital pathology; whole-slide image; WSI; image annotation; virtual slide; virtual magnification; histopathology; cancer diagnosis; cancer grading; genotype-

phenotype association; medical image analysis; clear cell renal cell carcinoma; kidney cancer

TYPE OF DISSERTATION: Theoretical Research; Application Research

目 录

CONTENTS

1 Introduction

Cancer diagnosis requires at least a set of histopathological or tissue slides where tissue samples are acquired by biopsies. The tissue slides are prepared and analyzed by pathologists. This tedious and repetitive task, together with more demands from growing number of patients, keeps pathologists occupied, preventing them to perform more meaningful task such as development of new diagnosis methods and closer examination of rare cases. Maturation of statistical machine methods fueled with flooding histopathological data creates a unique opportunity for clinicians, pathologists, and data scientists to collaborate and establish new systems that can assist medical personnel both in clinical and pathological routine to ensure that all cancer patients with different ethnicity, race, and geological origin could receive the best of care. Improvements may derive from more rapid, accurate, or comprehensive diagnosis, more time spent on rare cases by pathologists, expert systems working as pathologists in rural areas, and so on. The mentioned systems can be developed based on much needed biomedical infrastructure to support various type of data and its meaning in pathology, and even in clinics. The future of digital pathology would not be made made of machine entirely, but human with machine working together [1] to battle the deadly disease such as cancer.

1.1 Background and significance

Since cancer is one of a genetic disease, analysis of genetic materials will lead to the diagnosis of cancer. Along the way from genetic alterations to manifestation of cancer symptoms lies many steps which medical tests could unveil the hidden condition within each patient. This mechanism will be discussed in the following sections along with introduction to histopathological slide and analysis which is crucial component for all cancer diagnosis and treatments.

1.1.1 General concepts about cancer

Morbidity rate of cancer will only increase in the future with the current stage of global development in various perspective. In the United States alone, it is projected that there will be approximately 1,762,450 new cancer case which is the equivalent of more than 4,800 new cases each day in 2019 and about third of the number will die from cancer [2]. More people will develop the disease because of longer life expectancy, deteriorating environments in developing countries, food and air contaminations, and so on. Generally, the development of cancer is caused by errors in translation or transcription of genetic materials that cannot be corrected by natural mechanisms. Those cells with damaged DNA beyond repair are the root origin of cancer. Factors contributing the mentioned phenomenon could be, but not limited to, inherited genetic materials, longer lifespan, living conditions—environmental effects, un-

healthy behaviors—smoking or being exposed to radiations. With the mentioned factors, it can be predicted the number of cancer patients will grow in the coming decades.

Cancer mortality rate, on the other hand, can be limited thank to new screening, diagnosis, and treatment procedure that are more comprehensive and effective in distinguishing cancer patients from the general population and treat them. As a result of previous development in cancer screening and diagnosis, mass screening procedure such as prostate exam of male aged more than 40 or mammography have proven to be very effective and have prevented many cases of mortalities. Some well-established treatment procedures are effective to some oncological types or sub-types such as tissue resection, hormone treatments, nephrectomy, chemotherapy, and radiotherapy. A number of emerging treatment methods are being tested [3] e.g. immunotherapy and personalized approaches. In short, while there is an increasing number of cancer patients, there are more means of treatment and more opportunities to fully recover from the disease as well.

Cancer is known to be a genetic disease, altering the genetic materials so that tumorous clump of cells—either benign or malignant—develop into serious condition. Along the way, the disease can be detected at many levels. Earlier detection of the disease will make it easier to treat and is the goal of cancer detection. Characteristic of cancer can be recognized at different levels: genotype, phenotype, and clinics, and useful information can be extracted from each level at different time of life (or development of cancer) as illustrated in Figure 1-1. The genotypic data may include DNA or RNA sequencing, SNP-based platforms, Array-based DNA methylation sequencing, or Reverse-phase protein array, while phenotypic data may include tissue slides or radiology scans. From the illustration, it is feasible to predict the tendency if one is going to have cancer or not since the birth time by analyzing genes. Later in life, once the cancer has been manifested, early detection using at the phenotypic level can be done. Nevertheless, most of cancer symptoms will surface once the disease has been developed for a longer period of time, thus making it harder to treat.

As mentioned, there are many lines of defense before deaths from cancer will occur. However, there are many steps that the medical procedures could have gone wrong as well. Currently, cancer diagnosis is the crucial step that clinicians rely on to decide if particular patients have the disease or not, and decide the following action based on the extent of the disease diagnosed at this step of the clinical routine. To confirm the diagnosis, a lot of information and many experts are involved. At this point, clinical or medical records about the symptoms medical history of the patients is on hands, encouraging clinicians to request further specialized tests such as radiology scans for appropriate anatomical part of the body. Finally, to truly confirm the disease diagnosis, examination of tissue section is needed. This is done with an invasive procedure called tissue biopsy to cut samples of tissue from patients for closer examination with light microscope by pathologists. Since the procedure is invasive, clinicians must

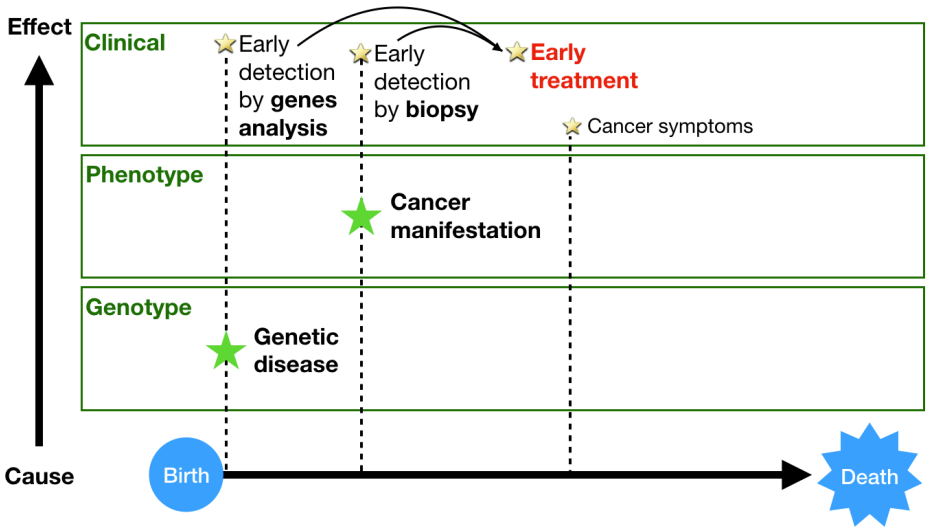


Figure 1-1 The development of cancer at different information level and time of life.

be sure enough to request the test. Histopathological slides acquired from the procedure retain valuable information about the extent of the disease, if there is any. In this thesis, we will discuss about the application of artificial intelligence in histopathological analysis because it could improve the quality of patient's life with more accurate diagnosis with richer information extracted from tissue biopsy.

Other than cancer preventions and treatments, after treatments have been administered for patients, following up the results is needed, this requires additional diagnosis to be made. Some also need maintenance treatments to keep the system cancer-free since some severe cases are most likely to have relapses. These aspects will not be discussed in this thesis.

1.1.2 Histopathological image and analysis

At the center of pathology lies histopathology which is the analysis of tissue samples as a whole. Manual histopathological image analysis by pathologists is the current gold standard for diagnosing cancer patients. Prognosis and treatment planning also heavily rely on the information from pathology analysis, often in the form of pathology report. Needless to say, pathology is a crucial part of any clinical routine involving cancer. Vast amount of information hidden within the patterns of histology samples was conventionally extracted by pathologists trained to interpret the meaning of images they saw using light microscopes.

Within the past decade, increasing availability of medical image, especially histopathological image—namely digital slides enable the accumulation of histopathological data in the digital form. This pave ways for computer algorithms access this type of data in large quantity for the first time. Together with cost reduction in digital storage and transmission, digital slides are capable of being mobile like never before. There have been many approaches to digitized

glass slides. Popular ones are connecting CCD cameras with microscopes, capturing the tissue image one diagnostic area at a time and using whole-slide scanners to scan the whole slide. Whole-slide scanners can capture the content of the whole slide and save the image file known as whole-slide image (WSI). Since the introduction of this new type of data, many researchers have tried exploiting it and come up with many new ideas assisting pathologists in diagnosing the tissue slides.

Reducing the workload of pathologists is important since they can move on and do more important things. For example, analyze more complex patterns or disease. In other words, automated systems will replace repetitive routines. At the same time, they will also solve inter-rater disagreement in tissue slide diagnosis which is high in conventional microscopy. Using automated systems for analyzing digital slides is a promising way to overcome the overwhelming need to manually interpret the tissue slides. However, more mature bioinformatic infrastructure to support this kind of system is much needed.

Preliminary works, findings, and proof of concept relating to automated digital slide analysis have confirmed that the approach is feasible. Extending those works needs better biomedical infrastructures or platforms. Factors limiting those experimental works in the real-world application is need for parameter tunings [4] because of high variations in histopathological images and non-standardized staining methods, result in a lot of noises in final scanned images. There are no international, or sometimes national, standard for pathological laboratories around the world to follow. Also, there are no open and standardized datasets for most cases of cancer and the open ones are not comprehensive enough. Most success in automated systems are in the area of deep learning where a huge amount of referencing data is needed to train the artificial neural networks so that it can effectively localize, segment, or classify histologic slides. The medical data is often not made publicly available from many reasons: patients' confidentiality, cost of acquiring the data, cost of annotation. Most of experimental tools are not commercial-ready, thus it is very hard for end-user like pathologists and biologists to use.

1.2 Data acquired from cancer screening, diagnosis, and treatment

The ultimate goal of a medical system is to effectively successfully detect patients with disease as early as possible which will make it easier to diagnose and treat. Timeline of the entire process is often consisting of three steps: screening, diagnosis, and treatment. In screening, the goal is to apply the procedure to the mass public, thus the procedure should be quick and simple with low cost and minimal burden to patients. Good screening procedure will result in early detection of the disease and low mortality rate. The diagnosis is close to screening, this is the step where the patients could get the confirmation if they have the disease or not, and the extent of the disease they are having. Sometime, the prognosis is made with the diagnosis to estimate the survival outcome of the patient. The treatment is administered based on the

diagnosis. There are different type of treatments and each of them are effective to only some cancer stage. Thus, accurate diagnosis should be carried out to save the patient from burden of the treatment.

1.2.1 Cancer screening

Cancer is often screened based on the risk according to the age group, demographic, and genetic inheritance (family history). At the present, only a few types of cancer benefit from analyzing data acquired during this process such as mammography to look for breast cancer and MRI brain scan to look for tumors. Medically, symptoms founded at this step is often unreliable and further tests will be carried out before the diagnosis can be made. The screening method is also noninvasive since it should not introduce unnecessary risks to the patients. There are some procedures that cannot yet be digitized such as examination for prostate or self-examination of breast.

Radiology scans is often used in cancer screening since these scans are noninvasive. Expertise needed to interpret the scans is also being assisted by computerized system. The development of fully automated systems is underway. The data used in cancer screening can only provide insights at the organ-level to screen for potential patients. It cannot clearly confirm the presence of the disease.

1.2.2 Cancer diagnosis

A lot of efforts have been focusing on automating cancer diagnosis over the past decade because this step is a repetitive, time consuming, error prone in term of disagreements between different raters, and it is predicted that there will only be more and more demand to diagnose more patients since human is having a longer life-expectancy thus more likely to have cancer. Conversely, it takes the same or more resource to train experts to recognize the disease and perform the diagnosis. It is clear that the medical community cannot keep up with this trend, and automated system or computer-aided diagnosis (CAD) is urgently needed to reduce the workload. Diagnosis is crucial to the clinical workflow since clinicians rely on the diagnosis to know the presence and extent of the disease.

Biopsy is the main procedure that doctors use for cancer diagnosis. Performing the biopsy may result in different type of data for example tissue slides stained with different kind of chemicals, gene sequencing, and proteomic analysis. Diagnostic tissue slides or histopathological slides is the gold-standard where pathologists will examine several slides from one patient and make the diagnosis. It takes years for pathologist in training to effectively recognize the patterns in the slides. The diagnosis on some complicated slides must be finalized by senior pathologists. It may be concluded that, currently, the heart of cancer diagnosis is that the analy-

sis of histopathological slides. Only recently that these histopathological slides were digitized using whole-slide scanners and produce an extremely large image, whole-slide image. This movement has enabled the digital processing and analysis of such type of data.

1.2.3 Cancer treatment

Different kind of treatment will be administered to the patients based on the extent of the disease. Patients with low-grade and localized tumor tends to receive the treatment where the cancerous site was surgically removed or resected so that the organ became disease-free. Resection will be made alone based on the premise that the cancerous cells have not been spreading to other part of the body. Patients with early stage of the cancer are likely to be treated this way. Survival rate of the patient with low-grade tumor or early-stage cancer is high. In the case of high-grade tumor or when it is assessed that the cancerous cells have metastasized, adjuvant therapy will be administered as the second treatment such as chemo-therapy, radiation therapy, hormone therapy, or other targeted treatments. These additional therapies are given to deal with cancerous cells outside the primary site. Deciding the correct type of treatment is crucial in successful cancer treatment, reducing burden from the treatment and introducing minimal risks.

1.3 Milestones of digital pathology

Currently, whole-slide scanner which is a crucial part of producing digital slides is not a standard equipment in major hospitals. The United States Food and Drug Administration (US-FDA) has just issued a recommendation document for this type of machine in 2017 [5]. Meanwhile, there are several questions and concerns about the use of whole-slide image system such as its validity compare to glass slides, the imaging pathway where each image pixels are being deliver from tissue sample to computer displays, etc. Mainly, whole-slide scanner, at the moment, are being used for research purposes and not for diagnosis purposes.

1.3.1 Large-scale repositories of digital slides

In 1999, [6] have described the idea of digitizing the tissue slide, paving ways for newer generations of whole-slide scanners to be developed. The two major milestones that ignite the research in histopathological analysis are two US government funded projects called The Cancer Genome Atlas Project (TCGA) [7] in 2005 and Gene-Tissue Expression Project [8] in 2013. TCGA collects tissue samples from individuals that have been diagnosed with various type of cancer. It accumulates many types of data including RNA sequencing, MicroRNA sequencing, DNA sequencing, SNP-based platforms, Array-based DNA methylation sequencing, Reverse-phase protein array (RPPA), and digital tissue slides of the diseased organ. While

all cases in TCGA are made from individuals with cancer, GTEx cases are obtained from deceased healthy individuals and tissue samples are taken from several anatomical sites. These are two first large-scale repositories that are freely available to the public and analyses should go on for years to come.

1.3.2 Tools for whole-slide image

There are a number of tools available for visualizing and primitive annotating WSIs. The fundamental tool for reading WSI files is OpenSlide [9] developed by a team at Carnegie Mellon University in 2013. It is a library for reading WSI files that can work dynamically to efficiently read multi-scale mage and is embedded within all software that has been developed later. OpenSlide has provided mechanisms for researchers to read WSIs in all proprietary formats which in a way broken down the barrier of proprietary file formats. Since OpenSlide does not have any kind of GUI, two more GUI tools are developed called Automated Slides Analysis Platform (ASAP) [10] and QuPath [11]. The two software can read and annotate WSIs and they are open source software. Together with the advent of OpenSlide, Gutman D. et al [12] have developed a web-based digital slide archive (CDSA) where the WSIs can be viewed via a web browser. The system is made possible by OpenSlide and JavaScript-based multi-scale image viewer called OpenSeadragon [13]. An extension to popular tool such as ImageJ [14] called SlideJ [15] was made so that it could support to visualize and analyze WSIs.

1.3.3 Competitions related to digital slide analysis

Within the past 5 years, there are a number of competitions involving analyses of WSI: Camelyon 16' and 17' [16], TUPAC [17], ICIAR BACH [18], and so on. Most of them are trying to promote the development of artificial intelligence methods to localize tumorous area of the tissue or assess the extent of the disease for each case. The competitions are only made on breast cancer, while some are deriving digital slides from TCGA repository. It is clear that extensions of competitions for other oncological types are needed.

1.3.4 Artificial intelligent systems in digital pathology

Since digital slide repositories were made available to the public, there are many implementations of statistical machine learning on the data repositories. Recently, deep learning algorithms are used to classify oncological sub-types and extent of the disease [19–27]. There are also developments of quantitative or computable phenotypic characteristics e.g. radiomics [28] in radiology images which is possible to apply the idea to histopathology. Artificial intelligent component can be added to conventional microscope with the help of maturing augmented reality (AR) technology assisting pathologists in localizing potential tissue area that needs spe-

cial attention in real-time [29]. Ultimately, a complete biomedical system to interpret particular patient from all aspects is expected and it needs association studies to computationally link different data modalities that have been traditionally biologically related since the beginning [30, 31].

1.4 Outline of this thesis

In general, this work tries to bridge two gaps. First is the biological gap between the cause of cancer, currently believed to be oncological genes, and its expressions that can be considered as patterns in histology samples. Acquiring detailed annotation on tissue samples and utilizing image analysis techniques will help deepening the understanding of underlying cancer mechanism. Second is the technological gap between conventional and digital pathology. This will help us to transfer invaluable knowledge in pathology to machine readable bioinformatic infrastructure, waiting to be further analyze in the future with emerging techniques.

This thesis is divided into three main sections, building toward the implementation artificial intelligence techniques in pathology to enhance cancer screening, diagnosis, and treatment. The development of a new software platform is needed in response to the lack of comprehensive tools to work with histopathological image: especially whole-slide image. Requirements to create a good tool for this purpose are different from general images and will be discussed in chapter 2.

As a demonstration, the proposed tool in the second chapter will be used to annotate a set of diagnostic WSI acquired from TCGA-KIRC project creating multi-expert annotation according to WHO/ISUP grading system. The annotation is done digitally, utilizing various functions of the proposed software platform.

In chapter 4, different approaches that researchers are using were discussed and a new approach is proposed which is analyzing the histopathological image based on underlying biological understanding. A demonstration of analyzing based on such approach is made on low-grade renal cancer.

The conclusions and suggestions are summarized in chapter 5. In this chapter, we investigate the potential usage of artificial intelligence in histopathology. Future of digital pathology and needed preparation are discussed here. Digital pathology is an emerging area of research, thus there has not been many works focusing in this area.

2 Framework for annotating large-scale histopathological image dataset

Large-scale histopathological images hold rich information about the microenvironment of cancer which are crucial for interpreting the corresponding genotypes in omics data. Pathology is one of the last medical specialties to be digitized [9] there has been limited software and tools that helps pathologists manage and analyze extra-large digital scans of glass slides—namely, whole-slide image (WSI). Furthermore, even fewer open-source software may be used for annotating and analyzing the WSIs. No existing software supports online multi-user annotation with detailed spatial resolution and semantic meaning. In this chapter, we present OpenHI—Open Histopathological Image, a publicly available open-source annotation framework for WSI. It can achieve pixel-level precise boundary and semantic annotation. It also supports online collaborative annotation. Eventually, the proposed framework will facilitate the large-scale histopathological image annotation and benefit machine-learning based phenotype extraction.

2.1 Background

2.1.1 Large-scale histopathological image dataset

Recent successes in applying statistical machine learning methods to solve real-world problems such as image recognition, speech recognition, gene expressions association, etc. are data-driven and thrive on large-scale data, namely big data. Most of them are complex systems that need tuning, thus require large amount of annotations to learn from [32]. It is expected that digital pathology can also benefit from this type of artificial intelligence, especially deep learning. One of the barriers obstructing the advance of such system is the limited size of the publicly available datasets.

To make a cancer diagnosis in various oncological types, pathologists would need to analyze one or multiple tissue samples from a patient's biopsy and decide the grading of each specific sample. This tedious procedure is manually carried out with a bright field light microscope [33]. In recent years, there has been multiple implementations of machine learning methods to enhance histopathology image analysis workflow by either assisting pathologists in image analysis or by establishing automated pipeline to analyze, by detecting and classifying the cancerous area, the WSI with high throughput and high precision to help reduce the workload of the pathologists [27, 34–36]. However, these works were derived and tested on datasets with finite size and variability due to limited availability of public datasets. With our proposed framework, rapid creation of such datasets could be accomplished, thus sophisticate

computational approach that holistically analyze the WSIs [37] may lead to a better grading decision.

One of the ingredients for curating annotated histopathological dataset is raw histopathological image. It is known that medical data is expensive to acquire. Collaborative efforts and government support are needed to establish one. In cancer research, raw large-scale histopathology image repositories have been made publicly available by TCGA [7], GTEx [38], and other projects. This is tremendous amount of data with different oncological type, but they lack annotations. In recent years, there are enriched WSI datasets used in different competitions [17, 39], however, they are limited by public availability, size, variability in cancer type, or spatially precise annotation. Enriching the images with annotations could greatly benefit the scientific community.

2.1.2 Efficient annotation

Annotation of medical data is expensive since it demanded labor for manual data annotation from experienced medical personnel [27]. Measures must be taken to reduce the cost of biomedical data annotation. OpenHI was designed as a collaborative annotation tool, ensuring that the most efficient annotation is carried out by acquiring only the essential data from the annotators while maintaining high-quality annotation.

Collaboration between pathologists, as a source of expert knowledge, and data scientists, who will manage the acquired data, is necessary to complete large-scale histopathological image data annotation and cross-validate the quality, thus the annotation framework would need to minimize technical configuration at the annotator end (pathologist) and maximize configurability at the data scientists end. The ideal software for this application would minimize the annotators' effort to annotate the image while capturing high information granularity including high spatial resolution and standardized semantic meaning. In the meantime, it should support online collaboration as well.

2.1.3 Related work

An annotation software capable of precise visual annotation and semantic information enrichment is highly demanded. WSI files are hard to be read efficiently because WSIs cannot be saved in standard image format due to its unusually high resolution.

1) Existing standards

In response, different whole slide scanner venders have come up with their own proprietary standard and file format. The DICOM or Digital Imaging and Communications in Medicine standard is the only public and general standard [40] that has been proposed but has not been adopted or popularized. However, in 2018, a push for utilizing DICOM for digital

pathology has been demonstrated [41].

2) WSI tools

In 2013, [9] have introduced an open-source library to read the WSIs called OpenSlide which later become the only available vendor-neutral tool to read the WSIs to date. Many other WSI visualization and analysis software have adopted the library to create web-based application [12, 42], stand-alone software such as QuPath [11] and ASAP [10], and extension, SlideJ [15], to ImageJ [43]. Around the same time, OpenSeadragon (OSD) library [13] was introduced as a web-based viewer for high-resolution zoomable image written in JavaScript. It is capable of viewing the multi-scale images including WSIs. OSD is then used in web-based implementation of WSI viewer.

The web-based implementation of OpenSlide with OSD to help visualize the WSIs from TCGA project on the webpage could be seen in the US National Cancer Institute's Genomic Data Commons. It helps the users to visualize the WSIs without downloading the entire large WSI file, however, it lacks the functionality to modify, annotate, or analyze the WSIs. In 2017, QuPath [11] was introduced as a cross-platform stand-alone software. It is a tool to view WSIs on local machine and it is capable of accomplishing many tasks including basic annotation of the WSIs and locally segmenting the image with superpixel algorithms. The most detailed annotation method that QuPath can achieve is selecting multiple points in the image to form a polygon, this approach is good for manually mark a small number of regions for human references, it is not detailed enough for computers. Furthermore, the annotation data made in QuPath must be managed manually and it does not provide centralized system to manage the WSIs or annotation data.

No open-source collaborative annotation software specifically made for histopathological image is publicly available at the moment. Besides, such software should allow online-collaboration to achieve high annotation throughput and maximize the accessibility for the users since they do not have to download the entire dataset and install additional software. It should also be able to manage highly detailed annotation with region-specific semantic enrichment. The challenges will be addressed with our proposed framework.

2.2 Challenges in efficient histopathological image annotation

In this section, different approaches that can improve the annotation efficiency will be discussed. There are examples of efficient annotation in other fields such as natural image [44] where millions of images were annotated with thousands of labels. However, annotation of WSI has posed new challenges because of different file size, image dimension (as described in the introduction), and the type of annotation needed.

2.2.1 Crowdsourced annotation

Crowdsourcing annotation can accelerate the annotation while keeping high annotation quality with multi-expert data. Achieving crowdsourcing needs the software to run on web technology. The current internet infrastructure was not built to support the casual transfer of relatively large files like WSI files—the size is generally around several hundred megabytes. This problem is addressed by OpenHI by only transferring the part of the image that the annotator is viewing. The platform that supports crowdsourcing should be able to handle simultaneous access and annotation. OpenHI utilize relational database server, MySQL, to handle such traffic.

2.2.2 Insufficient medical personnel

The biomedical annotations are expensive to acquire. The acquisition workflow must be optimized, easy to follow, and standardized. Thus, most annotation result could be achieved with least amount of time spent by pathologists (medical personnel). The scarcity of experts is not the only factor contributing to this problem. Pathologists around the work do not use the same grading standard to grade the sample. Pathologists in some countries must grade their sample according to national standard provided by the authorities. Some hospitals are behind in updating grading standard for some type because of rapid updates and existing personnel were trained on old systems.

2.2.3 Region boundary delineation

Using freehand selection to draw on the image is not convenient and does not produce precise boundary selection since the hardware—mouse and keyboard—was not designed for this purpose. Some works use point-based selection [45, 46] to mark at the center of nuclei, but this method does not work in the situation where area containing several nuclei must be selected. Some existing annotation tools [10, 11] support bounding box-based and polygon-based annotation. Annotations based on bounding boxes have proven itself to be effective in natural image. It is straight forward for annotators and easy to store the annotation information, four points. However, this type of annotation is only good for object localization and not masking since the annotated area will include the background as well. Polygon-based annotation is better at outlining the boundary of the region but achieving a truly precise outline would need a lot of vertices.

2.3 Framework design

The proposed framework was designed to be implemented on a web server therefore it could be accessed simultaneously via a web-browser by multiple users. To minimize effort of acquiring most detailed annotations with precise sub-region boundaries and semantic enrichment, our framework pre-segment the image into semantically meaningful sub-regions as demonstrated in Figure 2-1 by a widely used graph-based image segmentation method called SLIC superpixel [47]. In this case, the annotators can quickly select the sub-regions by simply clicking or dragging mouse through them via the graphic user interface. Additionally, our framework has the ability to freely access any regions of WSI at will with different zooming level based on OpenSlide being utilized as the others did [10, 11, 42].

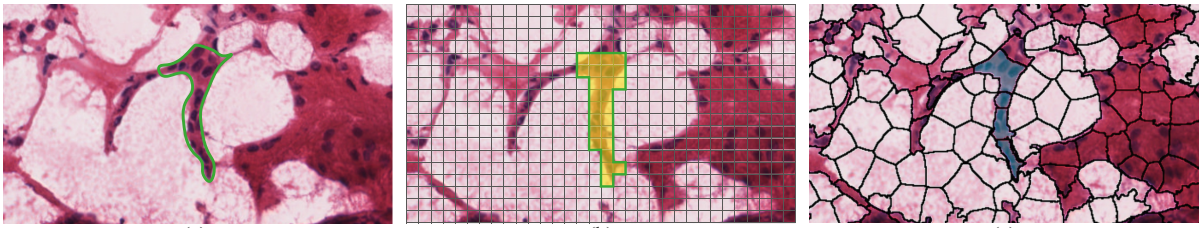


Figure 2-1 Comparison of how the intended region for selection (a) can be selected by a tiled-based (b) and superpixel-based (c) segmentation. Sub-region selection in (c) is more efficient since it needs less selection and achieve more accurate annotation.

The framework consists of three main components along with a MySQL server to store the annotation coordinates. The main components include image pre-processing of the WSIs, the web framework, and the GUI. Three types of data are stored in alongside the framework which is the WSI files, the sub-region boundary matrices, and the annotation coordinates. Figure 2-2 illustrates each component and the data flow between them.

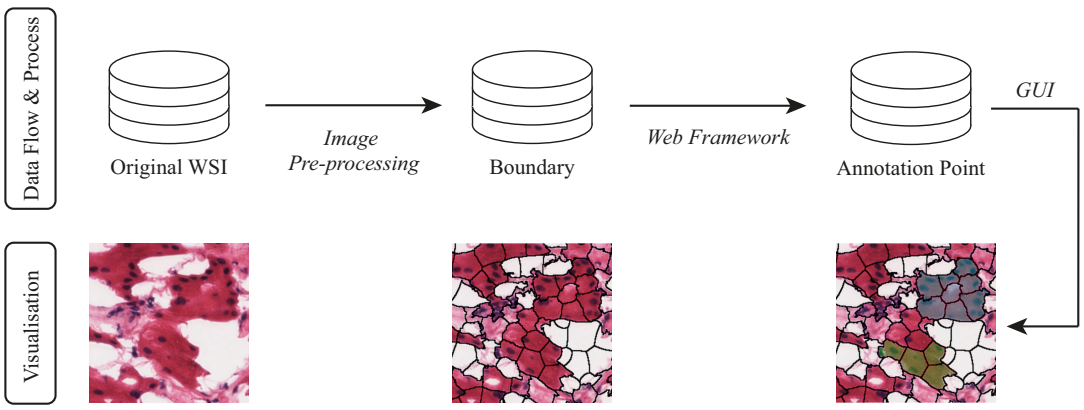


Figure 2-2 Structure of the framework with WSI data flow from original image to pre-segmentation and annotation.

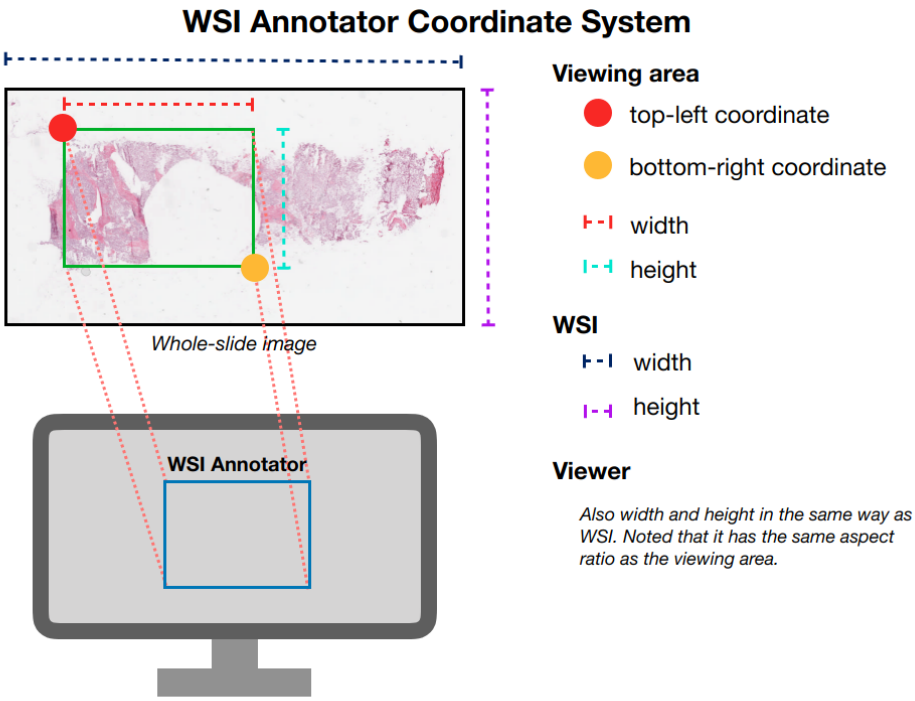


Figure 2-3 Denotations of terms needed for describing the viewing system of the proposed framework.

2.3.1 Pre-processing and boundary selection

A WSI is a very large 2-dimensional array of data. It is too large to be handled by most (and in some cases all) of conventional image formats. Furthermore, the conventional formats are currently not a vendor natural standard for such kind of data [30]. Conventional image segmentation technique has also faced a challenge in processing this kind of image as well. It is known that processing WSIs is memory demanding [15]. The SLIC Superpixel [47] segmentation on the WSIs also requires large amount of memory. Thus, it is reasonable to have memory-consuming processes deployed on a server with relatively larger amount of memory. To our knowledge, there is currently no practical implementation on superpixel with sufficient segmentation precision on whole WSIs on a scale of personal computer, however, there are efforts to alter superpixel algorithm for implementation with large images. Developing a robust and fast image segmentation method is still a challenge in digital pathology informatics [48]. By incorporating superpixel segmentation with the WSIs, we have established a new ROI selection method in digital pathology.

The lasting issue with superpixel algorithm in image segmentation is of making a decision about the number of final segmented sub-region which could lead to under- or over-segmentation. Tuning for good number of segments to avoid under- and over-segmentation in superpixel algorithm has been a challenge in utilizing the method. It is even more problematic

to choose one number for all images. To tackle this problem, we calculate the number of superpixeled sub-regions ($N_{superpixel}$) by specifying desired average sub-region size (P_{sub}), thus the size of the sub-region will be consistent throughout the annotation project as shown in Equation (2-1) where P_{total} is the WSI resolution. In practice, this number could range from around 6000 to 50 pixels/sub-region providing that the WSI has been scanned with 20x magnification lenses and it should increase or decrease with the magnification. The example of a portion of the image pre-segment with superpixel algorithm is shown in Figure 2-4.

$$N_{superpixel} = \frac{P_{total}}{P_{sub}} \tag{2-1}$$

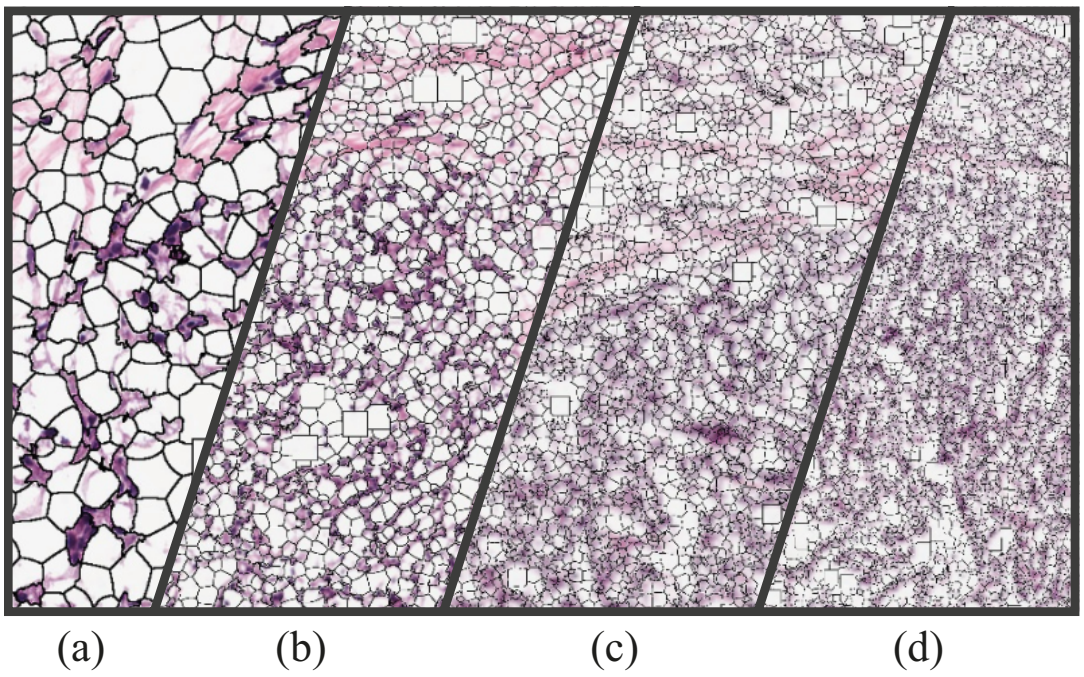


Figure 2-4 Parts of a WSI (a) is pre-segmented using SLIC superpixel algorithm at 6100 (b), 610 (c), and 60 (d) pixel/sub-region.

To avoid over-segmentation where an annotator must choose unnecessarily large number of sub-regions to establish one cancerous region or under-segmentation where a cancerous region is not cleanly divided from normal regions, our framework provides users an option to calibrate at multiple pre-segmentation levels and select the most suitable level for the annotation in specific area of the image. During the annotation session, the annotator could switch back and forth between different pre-segmentation level for a level that a cancerous region could be accurately separated at the same time with the consideration of the annotator's convenience and efficiency.

At the end of pre-processing, matrices containing sub-region boundary information are stored using a binary image format. This image with boundary information is large in dimensions but not in file size, therefore it is practical to store them as one continuous image where

it could be easily loaded into the memory when needed.

2.3.2 Server-end module

The second component of the framework is to interactively respond to the user requests in real-time. To lower the computational expenses, only the area of a WSI being viewed is processed. The processing includes the generation of sub-region boundary, visualization of the annotated sub-regions, and annotation coordinates recording and deletion.

As the use of OSD with OpenSlide has been demonstrated in [9] in 2013, we extend the usage for WSI annotation. Our framework has attained the capability of OSD and OpenSlide to view a WSI with smooth zooming and panning experience while adding the customized annotation capability to the framework. This allows the annotators to easily annotate any part of WSI they want. When annotating several sub-regions with the same grade, an annotator may select each individual sub-region by a mouse click. Rather than clicking the sub-regions one by one, the framework provides an option of clicking then pressing mouse over several sub-regions to do bulk annotation since the adjacent sub-regions tend to contain the same swamp of cancerous cells. Aside from conveniently choosing the sub-region, the GUI should provide some indicator to approximate the zooming power of current digital image state, this is also called virtual magnification.

1) Slide visualization

Several steps are needed to prepare the image before it can be sent to the viewer in the GUI according to the requested viewer size and viewing coordinate (see Figure 2-3). Three kind of image array must be prepared. First is a part of WSI (Figure 2-5 (a))). This can be acquired from the WSI file directly through OpenSlide. The required part of boundaries (Figure 2-5 (b)) array can be acquired from the cropping the boundary matrix loaded in the memory since the framework initialization. The boundaries are then mixed with the marked coordinates—this data was retrieved from MySQL database server—turning into the colored annotations (Figure 2-5 (c)). Creating one cluster of colored annotation requires a closed region and a representing point. The point acts as a seed point for flood-fill operation to mark the rest of the connected region, but not beyond the boundary.

The cropped boundary array is embedded into the WSI. The colored annotation is then alpha blended into the image. The default alpha value used is 0.4, 40% transparency. The final image is compiled as a single image, compressed in JPEG format and saved to disk. The link to the image is returned to the GUI and OSD will do a HTTP GET request to download the image, then display to the user. The overall generation process is illustrated in Figure 2-5.

2) Coordinate recording

The coordinated stored in MySQL database is stored as absolute coordinate according to

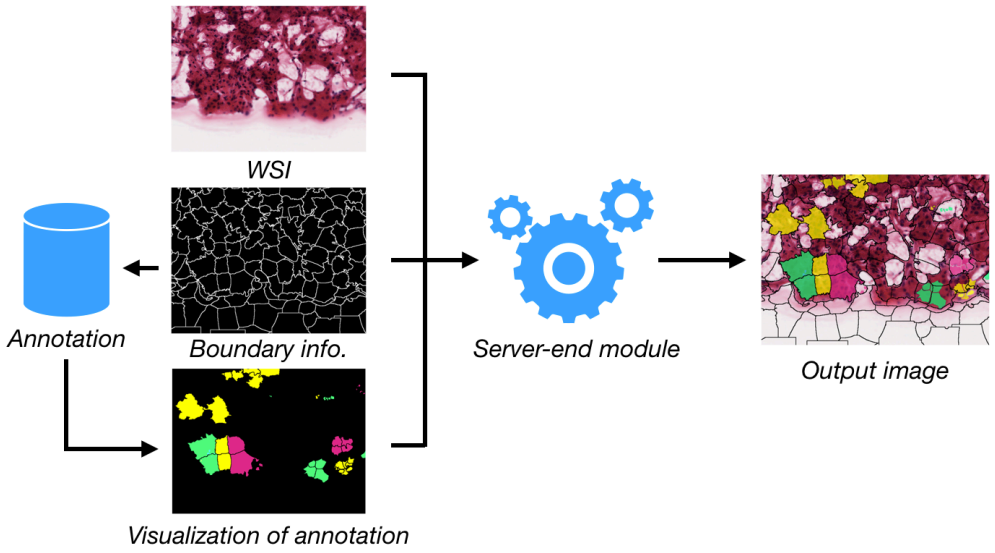


Figure 2-5 Slide visualization process in OpenHI where (b) is the cropped WSI, (c) is the cropped boundary image loaded from a pre-processed file, (d) is the existing annotation read from (a) MySQL database and colored using flood-fill operation, and (e) is combined view which is the final image and will be transferred to the viewer in the GUI.

each of WSI's dimension. For example, if the WSI has dimensions of 200 by 200, the top-left, center, and bottom-right coordinate will be (1,1), (100, 100), and (200, 200) respectively. However, the coordinate system used by OSD are relative coordinates referenced by floating-point numbers where the top-left, center, and bottom right of the viewer is (0.0), (0.5, 0.5), and (1,1), thus they must be converted from relative to absolute system according to the top-left coordinate of the current viewing area. Each point is stored with annotator ID, slide ID, corresponding PSLV, etc. the details are listed in Table 2-1.

3) Virtual magnification

Since the gradings of some cancer type such as WHO/ISUP kidney clear cell carcinoma annotation standard [49] (see Table2-1) require the annotators to take magnification power as a part of decision. This functionality is crucial for the grading system that rely on microscope magnification. It is also a worthy indicator for the pathologist who is new to digital pathology as well.

To accurately calculate virtual magnification, we need to understand that real magnification in the microscope is the combination of magnification from objective and eye piece lenses as in Equation (2-2). In virtual slides, the objective magnification M_{obj} is restrained by magnification factor of the objective lenses used during the scan which is specified in the metadata of the WSI file. The eye piece magnification M_{eye} is more complicated to calculate, three parameters are needed for the calculation including scanner sensor pixel size, monitor's pixel size, and distance between the monitor and the user. The scanner sensor pixel size is often not included in the metadata of the WSI file while magnification power is specified, thus it could

be referenced back to the scanner manufacturer about scanner's resolving power correspond to the specific magnification factor, and the sensor pixel size could be calculate the method by T. Sellaro et al. where there are calculation sample in their work [50] to obtain the objective magnification M_{obj}. Finally, the total virtual magnification M_{total} could be calculated by Equation (2-2).

$$M_{total} = M_{obj} \times M_{eye} \tag{2-2}$$

To avoid complications in our framework, we simplify the virtual magnification calculation and focus on three parameters which are the image pixel size (S_{WSI}), the monitor pixel size ($S_{monitor}$), and the digital zooming ($M_{digital}$). The image pixel size can be found in the metadata of the WSI file and will be given in the unit of length per pixel, often micron/pixel. This parameter reflects the real-world size of the scanning object which we use as an anchor or original size and the magnification is the zooming factor based on this size. The monitor size can be calculated by considering the screen resolution and the monitor size. OpenHI can automatically detect these setting with the standard JavaScript protocol. To calculate the magnification factor by incorporating the image pixel and screen size can be done by Equation (2-3). The digital zoom can be calculated by the size of the viewer shown in the GUI (S_{viewer}) and the size of images being viewed in the viewer ($S_{viewing}$) as shown in Equation (2-4). Finally, the virtual magnification in OpenHI can be calculated by Equation (2-5).

$$M_{WSI-monitor} = \frac{S_{monitor}}{S_{WSI}} \tag{2-3}$$

$$M_{digital} = \frac{S_{viewer}}{S_{viewing}} \tag{2-4}$$

$$M_{total} = M_{WSI-monitor} \times M_{digital} \tag{2-5}$$

2.3.3 Graphic user interface

The web-based GUI of our proposed framework is comprised of the main WSI viewer, virtual magnification indicator, and menus for annotation configuration as shown in Figure 2-6. The viewer, based-on OSD, enables an annotator to navigate around the WSI and the menu allows the user to change some configurations to make the annotation easier including the pre-segmentation level, tumor gradings, undo button, and option to show or hide sub-region boundaries and existing annotations. An alternative way to adjust these parameters is the usage of keyboard hotkeys that could speed up the annotation once the user gets familiar with the system. The GUI also has indicators which could assist the annotator to make a grading decision such as current virtual magnification factor, current resolving power, and the basic information about the tissue slide.

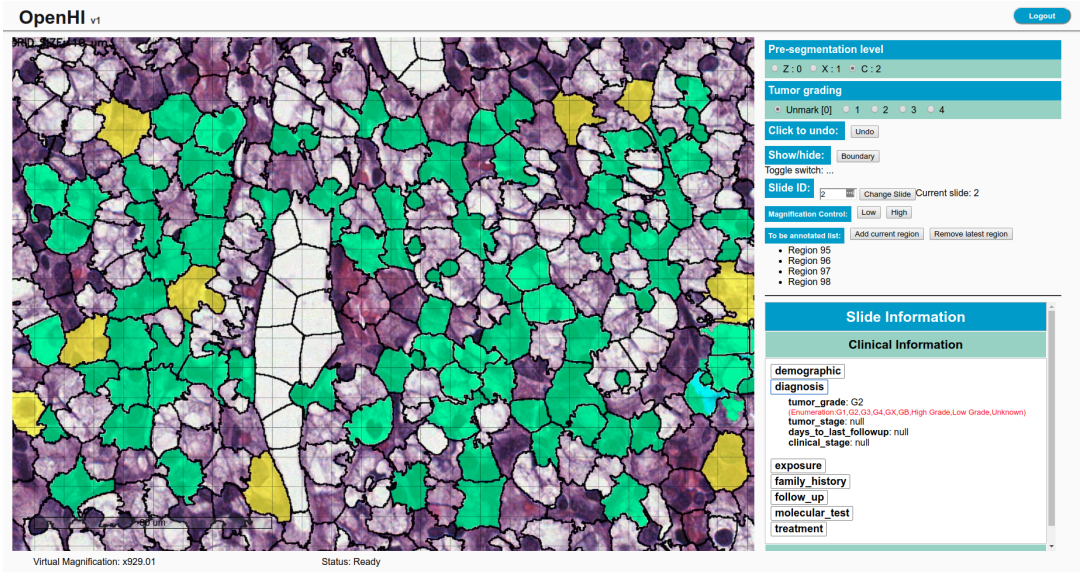


Figure 2-6 The graphic user interface of the framework with a sample image showing WSI viewer (left), virtual magnification indicator (bottom-left), control panel (top-right) containing tumor grading selector, pre-segmentation level selector, and slide information panel where accompanying information of a slide is shown.

2.3.4 Data model

To maximize the granularity of the annotation and keeping the utility of collected data. The framework will orient the data collection around the annotation coordinates by their properties including the exact chosen coordinate x and y, assigned tumor grading, pre-segmentation level, time of annotation, and the annotator's ID. The detail data model is described in 2-1. For full construction of the data model, see Appendix x.

All annotation coordinates are stored based on the level of WSI with the highest resolution to preserve the precision. Each of the coordinates is paired to the tumor grading and pre-defined boundary which the annotator has assigned during at the point of annotation.

To maximize the extendibility of this annotation software, it is designed to be highly customizable. Some parameters are designed for the annotators (pathologist) and can be switched during annotation processes. The other parameters are made for the data scientist so that they can achieve the sufficient annotation quality for further uses.

2.3.5 Software configuration

A successful annotation consists of a set of appropriate configurations to meet the needs of different annotation settings. We categorize the configuration into two group: annotator and initial configuration. The annotator configuration is options that can be easily changed by the end-user of the platform. This type of configuration can be changed back and forth

Table 2-1 Summary of data fields recorded in MySQL database.

Field name	MySQL datatype	Description
Point id	INT	ID of each annotation point/coordinate (x, y)
x-coordinate	INT	The coordinate x of the annotation point.
y-coordinate	INT	The coordinate y of the annotation point.
Annotation timestamp	DATETIME	Timestamp of the annotation action.
Grading	TINYINT	Grading of the cluster
PSLV	TINYINT	The PSLV that has been chosen by the annotator at the time of annotation.
Slide ID	SMALLINT	The unique slide ID indicating which slide the point belongs to.
Annotator ID	SMALLINT	The unique annotator ID specified by login information of the annotation session.
Region ID	INT	A unique ID assigned sequentially to each region.

and is part of the annotation. The user can change them using the control panel or keyboard hotkeys. Initial configuration must be set during the platform initialization and the parameters should be decided collaboratively by data scientist and pathologists, ensuring that high-quality annotation can be acquired.

2.4 Results

The scalability and extendibility of the proposed framework would make large-scale collaborative annotation of WSIs possible. It allows users to view WSI scans from multiple vendors and let them navigate around smoothly. Annotation with high information granularity can be accomplished. The proposed framework can ultimately be a foundation of crowed-sourcing WSI annotation platform. The enabled online real-time collaboration converges effort from pathologists to annotate and cross-validate large-scale images.

Table 2-2 The configurable parameters of OpenHI for extending to different image types and annotation settings.

Parameter	Time of configuration	Description
Pre-segmentation level	Annotation	This parameter determines the size of the sub-region in pre-processed
Number of grading level	Annotation	In different cancer type and grading standards. The level of grading is different.
Pre-segmentation sub-region pixel density	Experiment setup	This parameter could be mutual agreement between the annotator and the analyst before the pre-segmentation begins. Tuning of this parameter will allow efficient annotation process.
Grading tier	Experiment setup	Different cancer sub-type has different grading system. The annotation software could support different grading level.
Viewer size	Experiment setup	The viewer size could be fixed to a specific size or make it adaptive to the annotator's monitor screen size.
Fixed magnification zooming level	Experiment setup	OpenHI zooming can be fixed to different levels. Each level corresponds to specified down sampling factor of WSI.

2.4.1 Functionality

The proposed framework offers the freedom of choosing from pre-defined segmentation levels and switching during annotation. Meanwhile, the pre-defined segmentation levels are kept consistent on different images. For example, a sub-region in one image would contain approximately the same number of cells as another sub-region does in a different image, as long as the two sub-regions comply with the same pre-defined segmentation level. The consistency in the sub-region average size is achieved because the number of segments in superpixel segmentation was calculated by the desired sub-region size and the image size. The ability to support multiple pre-defined segmentation level which let the user to interactively switch

between them during the annotation session can increase annotation efficiency.

Figure 2-7 Comparison of different sub-region selection method where (a) is done by using polygons, (b) is region-by-region selection, and (c) demonstrate continuous selection across multiple sub-regions suing hold-and-drag method.

The proposed framework offers two methods of selecting annotation sub-regions which is a click to select individual sub-region or hold-and-drag over multiple sub-regions to select as illustrated in Figure 2-7 (b) and (c) respectively. The latter option is useful since a swamp of cancerous cells is likely to occupy several consecutive sub-regions. Selecting sub-regions in this manner could be relatively much faster than using relatively more mouse clicks to form an equivalently precise polygon as shown in Figure 2-7 (a). The comparison can be seen in Figure 2-7. Besides the easy sub-region selection, the user can also select gradings easily via the GUI next to the viewer as shown in Figure 2-7 or by hotkeys. Additionally, in the case that the annotator mistakenly selects the region, the framework can revert to previous step or the annotator can deselect some sub-regions.

2.4.2 Extendibility

In term of digital slide formats, since we utilize OpenSlide, the proposed framework can support various WSI formats from different scanner vendor including Aperio (.svs, .tif), Hamamatsu (.vms, .vmu, .ndpi), Leica (.scn), MIRAX (.mrxs), Philips (.tiff), Sakura (.svslide), Trestle (.tif), Ventana (.bif, .tif), and Generic tiled TIFF (.tif) [9].

Our software is free and open-source, it is available at https://gitlab.com/BioAI/OpenHI under GNU General Public License v3.0, therefore it can be modified to suit the need in different purposes. The framework is also compatible with general LAMP stack which is widely available on the could computing platforms or local server environment. For extending usability, the framework is fully documented and the docuemtnation is available at https://bioai.gitlab.io/OpenHI.

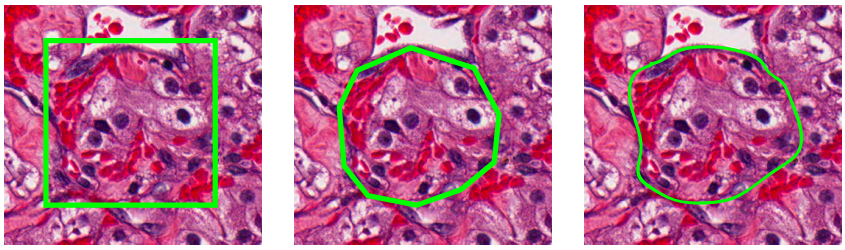
2.4.3 Demonstration

1) Testing environment

In software development and testing, we use WSIs directly downloaded from TCGA data repository [7]. Thus, the testing environment, WSI format used in our proposed framework is Aperio (.svs) file. The images are scanned with 20x magnification with resolving power of 0.5 micron/pixel [30]. In our sample set of data, the average resolution of WSIs is 920 megapixels with the maximum at 11,282 megapixels. The file contains three levels of multi-scale representation, and the average file size is 202 MB with the maximum of 2GB.

2) Performance

Annotation outcome and processing speed are the two main factors that contribute to the performance of the framework. The level of precision in annotation could be categorized into bounding boxes and circles, polygons, and pixel-level annotation as shown in Figure 2-8 (a-c) respectively. The mentioned annotation methods are ordered from greater simplicity with less precision to more complexity with more precision. Comparing OpenHI framework to other similar software, our proposed method can support pixel-level annotation precision while the other software [12-15] cannot. In comparison, OpenHI could significantly improve the annotation quality thus achieve better annotation outcome. Other than the annotation quality, the processing speed which contribute directly to the interactivity of the framework can reflect annotation efficiency. In our proposed method, the framework will consume more processing time when large area of the image is needed to be viewed. To demonstrate the robustness of the proposed method, it was tested based on the real-world usage as described later in this section.



(a) (b) (c)

Figure 2-8 Different level of annotation precision from different image annotation methods from the least precise to the most precise boundary including (a) bounding box, (b) polygon, and (c) pixel-level annotation.

The framework was tested on an Intel(R) Xeon(R) CPU E5-2650 v4 (2.20GHz), 1266 MHz with a total of 48 cores and 256 gigabytes of RAM. The current repository occupies 500 gigabytes of storage. The host operating system is Ubuntu 16.04 LTS. However, for single user, the minimum requirement for the host that we have tested with is Intel Core i7 (1.7GHz) with 2 cores and 8 gigabytes of RAM excluding image pre-processing due to memory limitation (see image pre-processing section).

The 250 MB WSI was used during the processing speed testing. The test was performed on a machine with single user requirement. The response time is varied by processing resolution

which is specified by the zooming level that the user has queried. For instance, if the user request to view a small area of the WSI or use high magnification, the processing resolution will be low. To view the image and sub-regions clearly, the user will need to magnify the WSI so that only less than 15 megapixels of resolution are needed to be processed. The examples of the viewing image on 800-by-460-pixel viewer at different processing resolution is illustrated in Figure 2-9. In most cases, comfortable viewing magnification is at 3 to 8 megapixels of processing resolution where it is suitable for annotation task. The response time is shown in Figure 2-10 where the average processing time will take around 300 ms with the maximum at 580 ms which is almost unnoticeable and responsive enough to perform annotation task efficiently.

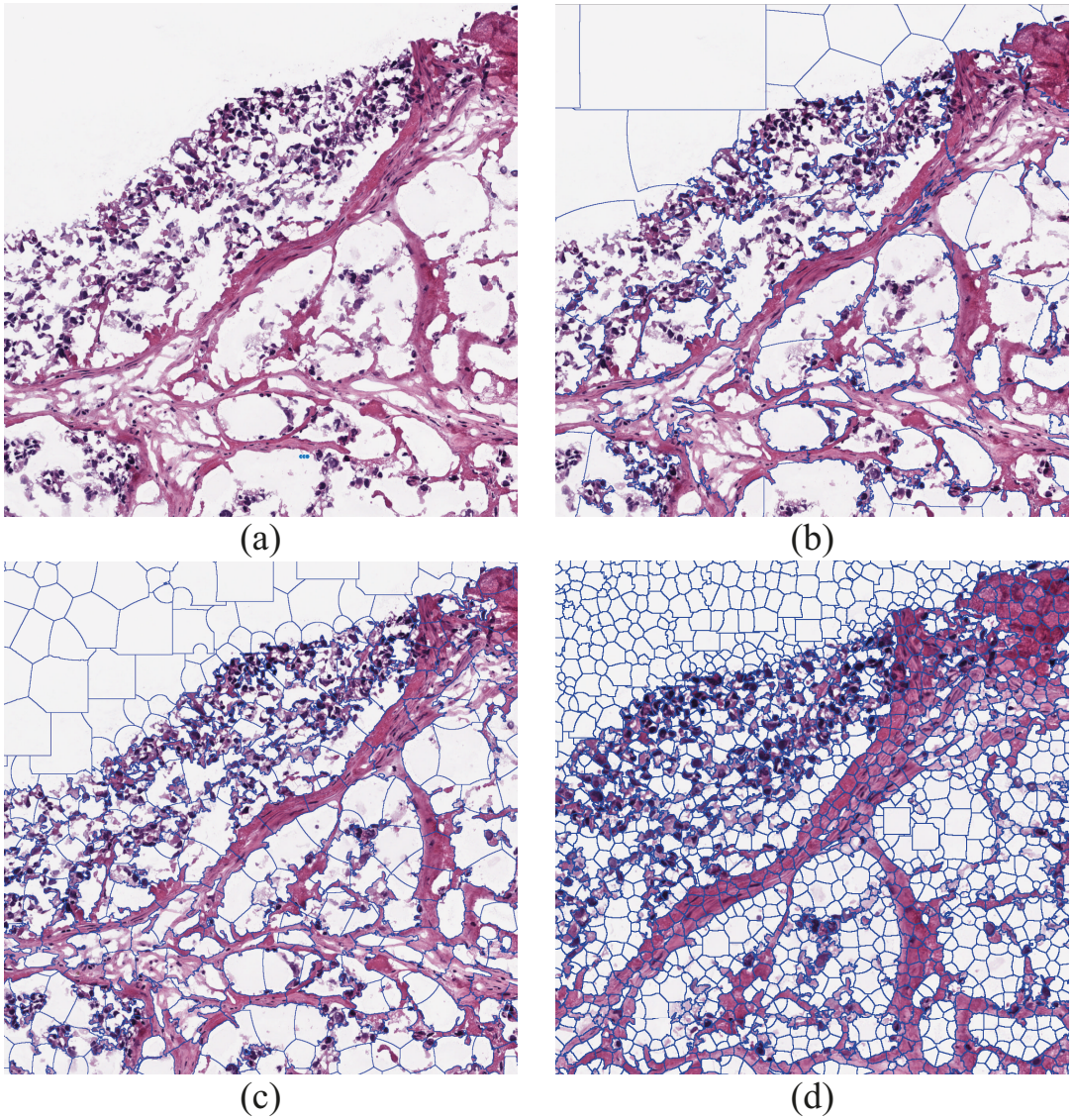


Figure 2-9 Example of viewing image in different processing solution ranging from 3.3 (a), 5.1 (b), 9.4 (c), and 15 (d) megapixel.

To conclude about the performance of the framework, in term of annotation outcome,

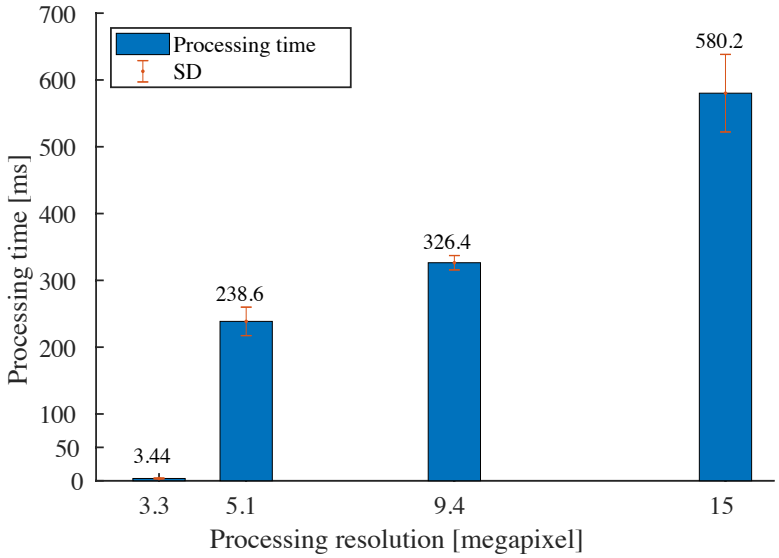


Figure 2-10 Processing time in different processing resolution ranging from 3.3 to 15 megapixel, corresponding to images (a) to (d) in Figure 2-9.

the framework can achieve superior annotation precision compare to other similar software while maintaining decent processing speed. Thus, OpenHI can be used to perform efficiently in large-scale annotation of WSIs.

2.5 Discussions

Several initial steps are taken in the development of OpenHI toward a histopathological image platform that can function fully as a part of hospital systems.

2.6 Open system for whole-slide imaging

Radiology imaging machine such as X-ray, CT, MRI, and other systems are offered or sold as a software and hardware solution—a complete system, but they are not being a fully closed system. In a sense, a microscope can be considered as a closed system and its components are being isolated into whole-slide image scanner and visualization tools. The US-FDA has approved the WSI imaging system as a whole in 2017, looking at the image pathway or how each pixel of the image are being taken and deliver to computer displays. It can be inferred that the quality of WSI viewed on third-party computer displays cannot be guaranteed. This obstruct the free data sharing of WSI. Several proposals made in section 3.2 about simulating microscope magnification solve parts of the problem. The open source nature of OpenHI platform also makes every pixel shown on the web browser tracible to its source. To enable crowd sourcing, telemedicine, and CAD in WSI, open imaging systems is much needed.

Generally, pixels in WSI files produced from whole-slide scanner originate from a light

source, passing through objective lenses and recorded with an image sensor array. There are many parameters during each step of the light pathway, and they could be between scanners from different manufacturer. OpenHI calculates virtual magnification factor based on image pixel size, 0.25 micron/pixel in our testing dataset. This parameter was not provided in the metadata of every WSI file format. A recent regulatory guideline for whole-slide imaging system [51] by the US-FDA does not specify that WSI files should provide this parameter. Instead, it focuses on a closed whole-slide imaging system spanning from scanners to computer displays which is not good for data interoperability. Furthermore, simulating virtual magnification should bear in mind that image presented on computer displays and perceived by annotators may be different.

Currently, OpenHI is based on web technology and crowd sourcing and it will hardly be a closed system since the platform needs to interact with several different devices, mainly computer displays. We need to study the effect of using different devices to do the slide interpretation while not strictly controlling the specification of the hardware used. Ensuring the image pathway stays intact from the beginning (WSI file) to the destination (each pixel lighting up on computer displays) is complicated since there are many involving factors. In OpenHI, the image pathway is controlled by the down sampling factor and the viewer size.

2.6.1 Required processing throughput

The baseline of the processing throughput should be considered the image throughput which conventional microscopes can deliver. Earlier, we discussed about the processing resolution where OpenHI can process up to 15 megapixels while maintaining a decent response time. To translate the conventional sense of analogue to digital resolution, we rely on two concepts including viewing area and resolving power. The resolving power can be considered as the smallest visual unit that someone can see using the microscope. Compensating the loss according to Nyquist sampling theorem, the double of resolving power should be equivalent to digital resolution. The resolving power in bright field microscope relies on the objective lenses ranging from around 3 to 0.25 micron. With objective lenses of 100x magnification, the resolving power is around 0.25 micron. Combined with a typical eye piece lenses with the field number of 22, the field of view would have a diameter of 550 micron. This would yield the digital processing resolution of around 8 megapixels.

Noted that current 1920 by 1080 pixels computer display, known commercially as “Full HD” has around 2 megapixels. Most standard GPUs can support two Full HD display, thus having the processing power of 4 megapixels at the refresh rate of around 60 frame per second (fps) or approximately 16 ms response time. Sufficing to say that current computer’s hardware does not have the capacity to fully process and deliver image with quality as high as what pathologists can see using microscopes and naked eyes.

2.6.2 Processing bottlenecks

According to the Performance section, the response time increases when the framework needs to process larger processing area. This behavior has many contributing factors. From our observation, we narrow the problem down to WSI loading and boundary image down sampling operations. Besides the response time, the framework startup time is also long. This is mostly caused by reading the boundary image into the memory.

1) WSI loading

While OpenSlide can efficiently operate on WSI files written in hierarchical image structure with multiple pre-process down-sampled layers—typically 3 layers in our work, it still takes time to perform the down sampling operation according to the input down sampling factor so that image in the viewing area can fit the required viewer size and be able to work with other layers of images in OpenHI, boundary and colored annotation for instance. This problem arises often when the access is needed at higher resolution. Upgrading to better hardware with faster disk accessing speed may be able to resolve this problem.

2) Boundary image down sampling

Similar to WSI loading, the down sampling operation of a boundary image also increase the response time of the system. Since the coloring of annotated c-cluster must always be operated at the highest resolution to avoid errors in flood-fill operation caused by interpolation, boundary image does not have any pre-processed smaller versions. Requesting large viewing area translates to cropping large area of the boundary image and down sample it and it takes time to do so. This problem was partially address in OpenHI by disabling an option the view the boundary and existing annotations at lower level of zooming—when the magnification power is not high enough.

3) Boundary image loading during framework startup

It takes several seconds for the framework to startup. This is mostly due to the read operation of several boundary images since one WSI has many corresponding PSLV and thus many boundary images. This issue was not dealt with in the current version of OpenHI, but the remedy may be faster disk read and write speed like in WSI loading.

2.6.3 Comparison of digital resolution in WSI and resolvable unit in microscope

Practicing pathology, especially histopathology heavily relies on the operation of microscope. All pathologists must be trained to use and understand the tool well. Making the transition from glass slides to virtual slides, comparison should be made clear so that users of the digital system can interpret image as well as they can when using conventional microscope.

In conventional microscope, the resolving power was tied to the characteristic of different objective lenses. This parameter is not the one that pathologists usually rely on; they rely more on magnification. The resolving power of every lenses can be calculated using Rayleigh's resolution limit formula so that, if needed, the users can calculate the resolving power they are using. This parameter provides insights to how small is the smallest entity that they can currently observe.

In virtual microscopy, tracing back to identify the current resolving power of image displayed is possible. According to Nyquist sampling theorem, it can be calculated from tissue the half of surface area display in a certain pixel. For example, the displayed pixel has the size of 0.5 micron can be translated to the resolving power of 1 micron.

2.6.4 Lack of inter-rater reliability tests and multi-expert annotation merging

While OpenHI do support multiple user during the annotation on the same tissue slide, interpretation of multi-expert annotation has not been included in the framework yet. Methods to effectively combine multi-expert annotation should be investigated. There are two aspects consolidate multiple annotations: spatial and grading degree. Similarly, methods to validate inter-rater reliability should be established as well. Using collaborative platform as an infrastructure, real-time validation should be possible. Moreover, discordance between pathologists do happen even in glass slides setting [33, 52]. Some are caused by pathologists' experiences and some do not have clinical consequences, being able to distinguish genuine pathological disagreements from the ones influenced by the annotation framework should be useful.

2.6.5 Methods to improve annotation efficiency

Further acceleration of annotation can be made by building on OpenHI using different approaches: provide machine-learning based suggestion, improve segmentation with nuclei localization, provide basic analysis the slide region, etc. Machine-learning based suggestion system is a promising approach to improve annotation speed since annotators would only need to confirm majority of annotations and correct some. But this approach may induce biases and affect annotation quality. The current pre-segmentation method could be improved by localizing areas with nuclei and only pre-segment those areas. It could save some efforts since annotators can only focus on areas with cells. Basic information about the viewing area could be provided such as nuclei density, nuclei size, or ratio of H&E color.

2.6.6 The distribution of OpenHI as free and open source software

Addressing the problem which is lack of neutral histopathological image platform with decent usability. OpenHI must be distributed as free and open source software (FOSS), making

it available to as many users as possible. Releasing the source code to the public available alone does not solve the entire problem since it would lack (1) usability, (2) documentation, and (3) extendibility. In this section, strategies and best practices to promote OpenHI to be FOSS is discussed including (1) use of central git repository, (2) methods for documentation, and (3) usage of automated tests and automated systems to run them.

1) Central public software repository and version control

First of all, the source code must be available to the public. But the code must be constantly improving to keep up with new demands and infrastructure updates. Modification by other volunteering developers should be possible. The solution to these requirements is central public repository with version control. OpenHI adopts git as a version control system and GitLab to host the source code. Dealing with legal implications, appropriate license can be applied to the repository. The distribution of OpenHI as a ready-to-use Python package is possible via Python Package Index (PyPI), then users can download and install the package using "pip" command.

2) Standardized documentation

Documentation of the software is important since it provides explanation on both how the software works and how to use the software. There are several available automatic documentation generation tools. In this work, we use a combination of several popular standardized tools and formats. The project structure is organized into modules consisting of classes and functions. Each function will have its own inline-documentation describing the objective of the function and its input(s) and output(s) parameters. Optional examples can be provided in the inline-documentation as well. The format of the inline-documentation is reStructured-Text (RST) (http://docutils.sourceforge.net/rst.html) which is compatible with python-autodoc module. It is capable of translating text-based content into well-organized and easy-to-read web-based document. An example of this is shown in Figure 2-11.

The engine of documentation used is called Sphinx which is a tool that can organize inline-documentation or Python docstring into sections with indexes. The index pages were also written in RST format. Sphinx tool can build the source files into multiple formats: HTML, PDF, doctree, etc. The HTML website can be uploaded and used as a reference of the software. Organizing the documentation in this approach yields many advantages: the documentation will be updated with the function, the need for arranging the documentation according to the code is low, it is standardized and can work with other tools like code suggestion, and so on.

3) Utilizing automated testing tools to assist further development and maintenance

Maintaining a working software is a lot of work. Utilizing code tests (or unit test) with automated testing tools will alleviate some workload and help to maintain the code quality. In OpenHI project, Python's "unittest" module is used to test functions. With GitLab's "CD/CI"

```python
def get_original(annotator_id: int, slide_id: int, part: int) -> np.ndarray:
    """Get original image based on annotator, slide, and annotation part

    :param int annotator_id: The annotator ID in the OpenHI system.
    :param int slide_id:
    :param int part:
    :return: Original image extracted from WSI.
    :rtype: ndarray

    :Image result example:
        .. figure:: _static/original.png
           :scale: 30%

        Raw original image extracted from WSI file.
    """
    pass
```

(a) Inline documentation

openhi.img.legacy.get_original(*annotator_id: int, slide_id: int, part: int*) → numpy.ndarray

Get original image based on annotator, slide, and annotation part

Parameters

- **annotator_id** (*int*) – The annotator ID in the OpenHI system.
- **slide_id** (*int*) –
- **part** (*int*) –

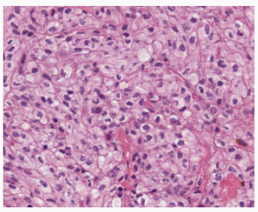
Returns

Original image extracted from WSI.

Return type

ndarray

Image result example



Raw original image extracted from WSI file.

(b) Generated documentation by Sphinx

Figure 2-11 Example of (a) text-based inline documentation written in reStructuredText and (b) output parsed with autodoc tool.

feature, the tests can be run automatically when changes are pushed to the repository. This way, maintainers will know almost instantly if there something is wrong with the new code.

3 Digital tissue slide annotation in clear cell renal cell carcinoma

Different oncology types require different diagnosis methods. They also have different degree of semantic meaning to each aspect of the annotation. To curate a dataset which captures all semantic meaning considered by pathologists during the routine diagnosis, the complete pathological routine must be studied throughout. In this thesis, the annotation of clear cell renal cell carcinoma (ccRCC) based on WHO/ISUP grading standard will be discussed.

The dataset used in this work is The Cancer Genome Atlas Kidney Renal Cell Carcinoma (TCGA-KIRC) which has 519 cases. About 80% of the patient in the dataset were diagnosed with grade 2 and 3 tumor. Each patient has one corresponding diagnosis tissue slide digitally scanned and saved in whole-slide image format with 40x magnification, 0.25 mpp resolution. The average size of the whole-slide image in this repository is 0.92 GB (0.46 SD).

3.1 Tissue slide assessment

3.1.1 Grading standard

To reasonably annotate the tissue image of ccRCC type, the widely adopted grading standard is chosen to be followed. There are several grading guidelines proposed for ccRCC over the decades. The most well recognized two are Fuhrman [53] and ISUP [49] grading guidelines proposed in 1982 and 2013 respectively. Both are 4-tier grading system based on nucleolar prominence in grade 1 to 3 and tissue architecture on grade 4. In 2013, the International Society of Urological Pathology (ISUP) has proposed an international grading system derived from the consensus decision of an international panel of pathologists which is simpler than and meant to replace Fuhrman's system since it disambiguates the grading criteria by discarding the redundant elements. Later, this grading system is adopted to be the international standard by the World Health Organization (WHO). The grading system is included in the World Health Organization (WHO) classification of urogenital tumors book (commonly called WHO's "blue book"). In this work, we thus choose to follow the WHO/ISUP grading system.

The WHO/ISUP grading system a 4-tier grading system. Its grading criteria is magnification dependent, thus, to strictly follow the guideline, the physical zoom of the microscope is required. The first three grades are based on the nucleolar prominence. The last and highest grade is based on the visible patterns embedded in the tissue sample. The fact that pathologists have to seek for the cell with highest grade implies a throughout search for the entire tissue slide. The grading standard is shown in Table 3-1.

In practice, searching for particular patterns was not done sequentially due to time constrain. Pathologists will examine the slide at the low power field and try to find the diagnostic

Table 3-1 WHO/ISUP grading standard

Grade	Description
Grade 1	Having inconspicuous or absent nucleoli at x400 magnification
Grade 2	Nucleoli should be distinctly visible at x400, but inconspicuous or invisible at x100 magnification
Grade 3	Nucleoli should be distinctly visible at x100 magnification
Grade 4	Tumors should encompass tumors with rhabdoid or sarcomatoid differentiation or those containing tumor giant cells or showing extreme nuclear pleomorphism with clumping of chromatin

regions. These regions of interest normally have high cell density. Pathologists will avoid the area that has necrosis, hemorrhage, edema, and area with high fibrosis since those areas do not have useful information that will contribute the diagnosis decision.

To make the decision for the slides which may have grade 1, 2, or 3, pathologists heavily rely on the microscope's magnification. Since the appearance of the nuclei is different at low and high-power field. The predominance of the nucleoli would have changed according to the magnification. The key difference between grade 1 and 2 where, at first, pathologist would examine certain diagnosis area at 100x magnification, if the nucleoli is inconspicuous or invisible, higher magnification will be used to confirm the diagnosis of grade 2. However, if the nucleoli are not distinctly visible at high magnification, the appropriate diagnosis should be grade 1.

3.1.2 Level of assessment

Tissue slide assessment can be divided into different levels: (1) patient, (2) tissue slide, (3) diagnostic area, and (4) individual cell. It is important to understand how different level of assessment affect the final diagnosis for a patient. During the cancer diagnosis, pathologists will assess several tissue slides to ultimately assign the overall tumor grade for each patient, this is regarded as patient-level grade. The patient-level grade is likely to be the result of several smaller decisions at the smaller level coming from one or more raters. Usually, several diagnosis slides will be made from one tissue sample or biopsy, the slide can have its own grade, and will be referred to as slide-level grade. Coming up with the slide-level grade requires a throughout examination of the slide. Pathologists usually will not examine the slide at every area since it is practically impossible to do so due to the time constrain. They will select certain areas, diagnostic area, to do the examination after looking through the slide briefly. At each diagnostic area, a diagnostic region-level grade can be made. Likewise, several diagnostic region-level grade will be used to consider the slide-level grade. Depending on the

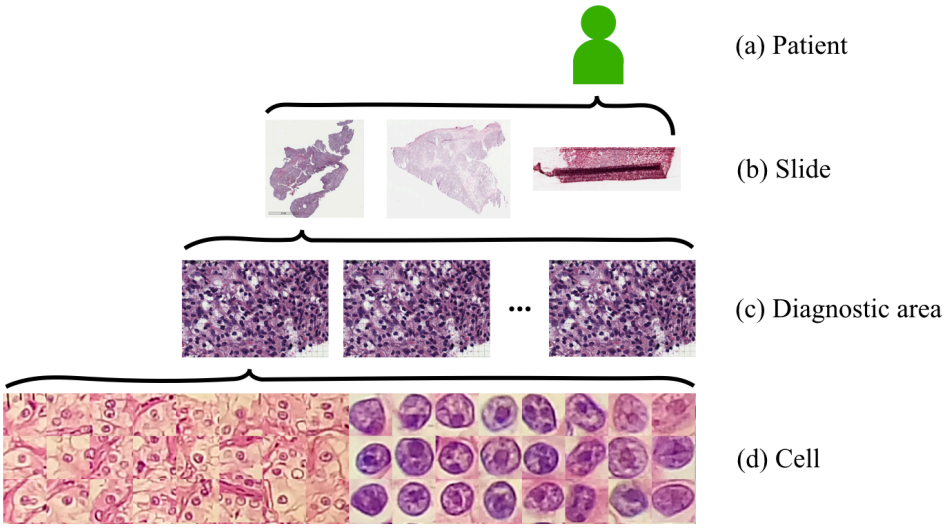


Figure 3-1 Illustration of how lower-level assessment affect the overall diagnosis starting from (d) cellular level, (c) diagnosis area-level, (b) slide-level, and (a) patient-level diagnosis.

grading guideline, average or the maximum grade may be used. In some grading system, including WHO/ISUP grading standard, the assessment can be made at the cellular level. The WHO/ISUP grading system is considered to be a nuclear-based system in in the lower grades (grade 1 to 3). Thus, grade could be assigned to individual nucleus.

3.2 Simulating microscope magnification using digital images

Accurately displaying the digital tissue image of 100x and 400x magnification to the pathologists on the computer screen is important especially when the diagnosis for ccRCC must be made since the WHO/ISUP grading system is magnification dependent. Existing software that can visualize whole-slide image such as QuPath [11] and ASAP [10] calculate the built-in magnification factor based on digital down and up sampling factor which is a totally different concept from microscope magnification. The awareness of the difference between digital and microscope magnification should be made in the digital pathology community. In 2013, [50] have addressed and discussed the factors that cause the appearance of the digital tissue image on different display to be different. In our opinion, the relationship between the magnification and resolution cannot be fully explained with current knowledge of how human eyes behave and the unknown relationship between the image which human eyes can observe and the perceived image. In this work, we utilize the concept of resolving power which is more straightforward and simpler to implement.

It should be noted that, the perceived image that pathologists can see in the microscope cannot be digitally displayed using today's typical equipment due to limited resolution, both at displays and scanners. There is a possibility that augmented reality technology may be able to reproduce the experience that pathologist has seen in the microscope, but studies for feasibility

is needed.

There are several factors which will influence the displayed image quality including image digitization, resolving of microscope (objective lenses), wavelength used, and image down sampling technique. In this section, we discuss how each factor affect the image reproduction and what are the optimized solutions that can be used to best reproduce the microscope experience on the digital displays to best simulate the physical microscope's magnification and help the annotators to come up with best judgment.

3.2.1 Magnification and resolving power

Adjusting microscope's magnification by changing the objective lenses will affect two factors which are the magnification itself and the resolving power. By definition, magnification factor reflects how big or small the object will appear to be. However, the definition for resolving power is different and defined by a microscope manufacturer (Olympus) as follow:

"The resolving power of an objective lens is measured by its ability to differentiate two lines or points in an object. The greater the resolving power, the smaller the minimum distance between two lines or points that can still be distinguished. The larger the N.A., the higher the resolving power." (Olympus [54])

Hence using high magnification alone to view the image is not enough to see the specimen clearly. The object that has been magnified beyond its resolving power will appear big but blurry, while the object that has been resolved with high resolving power lenses will appear more clearly.

Preserving the magnification from the slide digitization to image displayed to the pathologists on the computer displays is challenging since computer displays came in different sizes and resolutions. Also, the distance between the user and the display cannot be strictly controlled. Nevertheless, resolving power can be carefully calculated and the physically resolved entities can be deliver from the scanner to the computer displays. The idea is to define the smallest unit of a scan to be the same level as the resolving power. The minimum digital resolution (pixels) covering the area must be double the amount of resolving power unit to avoid loss in signal sampling according to Nyquist sampling theorem. Finally, each image pixel must have the corresponding display pixel to fully and independently display its color value.

For example, the object which has the dimension of 1-by-1 micron scanned using the lenses with resolving power of 0.5 micron would create four resolving units. To fully display these units on a computer display, 16 pixels are required.

3.2.2 The calculation of resolving power

The microscope's resolving power can be calculated by the Rayleigh resolution limit (R) [55] as in Equation (3-1).

$$R = \frac{0.61\lambda}{NA} \tag{3-1}$$

where λ is the wavelength used to visualize the sample and NA is numerical aperture value usually specified by the manufacturer of the objective lenses. Table 3-2 shows the typical parameters of microscope's objective lenses including the magnification factor and numerical aperture value.

Table 3-2 Typical bright field microscope's parameters used by pathologists.

Magnification	NA	Resolving power [μm]
4x	0.1	3.146
10x	0.25	1.366
40x	0.65	0.526
100x	1.25	0.273

To accurately calculate the resolving power, appropriate wavelength should be used. Generally, the wavelength of 530 nm is roughly used since it is the average of visible wavelength. More accurate wavelength can be interpret from the reported spectrometry of H&E stained tissue sample in [56] and the wavelength emitted by such type of specimen is around 530 to 580 nm. In our calculation, we take the full-width half maximum (FWHM) value which is 560 nm.

The resolving power is typically reported in the unit of micron in physical microscope. This parameter became micron per pixel (mpp) in the whole-slide scanners where a single pixel represents certain amount of physical area scanned by the scanner.

Using the mentioned parameters, we can calculate the Rayleigh resolution limit for 100x and 400x magnification used in WHO/ISUP grading system as 1.37 and 0.53 micron respectively. Consequently, the resolved entities in the image must be deliver to the annotator with no loss of quality. Taking Nyquist sampling theorem into consideration, the required resolution (P) of the digital slide needed to display the specimen correctly less than half of the resolving power which is 0.27 mpp (see Equation (3-2)). Thus, the resolution of WSI from TCGA-KIRC which is 0.25 mpp should be sufficient.

$$P = \frac{R}{2} \tag{3-2}$$

3.2.3 The calculation of down sampling factor based on resolving power

Image down sampling process reduces square root of pixels to a single pixel. Down sampling a digital image will produce the zoom out effect. For example, the image with the dimension of 2 by 2 (4 pixels) can be reduced to 1 by 1 with the down sampling factor (DSF) of 2. DSF used for our annotation can be calculated from Equation (3-3).

$$DSF = \frac{P}{S} \tag{3-3}$$

where S denotes pixel size of the WSI. The minimum DSF is 1. DSF lower than 1 indicates in sufficient digital image resolution.

The microscope magnifications specified by the WHO/ISUP grading system has encompass us to use the resolving power of 10x and 40x objective lenses. Those resolving power are 1.37 and 0.53 micron respectively. Using our approach to calculate the DSF required to process the image that will be displayed to the annotator, we concluded that the appropriate DSF for both low and high magnification are 2.91 and 1.12 respectively. In practice, however, we round the DSF of high magnification from 1.12 to 1 since so that the image displayed to the annotator will have original quality of the digital file and no interpolation would be used. The calculation details and other parameters used during the annotation is described in Table 3-3.

Table 3-3 Parameters used to set OpenHI low and high magnification.

	Source	OpenHI Low	OpenHI High
Microscope equivalent [total (objective + eye)]	WHO/ISUP grading standard	$100x(10x + 10x)$	$400x(40x + 10x)$
Wavelength	FWHM of H&E curve	560 nm	560 nm
Resolving power [μm]	Rayleigh resolution limit (R)	$0.65(0.56)/0.25 = 1.456$	$0.65(0.56)/0.65 = 0.56$
Required digital resolution [[μm]/pixel]	Nyquist resolution requirement (P) P = R/2	$1.456/2 = 0.728$	$0.52/2 = 0.28$
Available resolution	TCGA-KIRC WSI metadata	0.25 μm/pixel	0.25 μm/pixel
Selected resolution [μm]		0.728	0.25 (equal to original scanning resolution to avoid interpolation)
DSF	Equation (3-3)	2.91	1

3.2.4 Limitations of the resolving power-based approach

The development of resolving power-based approach, we have fixed some ambiguous factors that may be changed if the hardware specifications changes in the future. The requirements are as follow. (1) Human annotators must be able to see or resolve every pixel of the display. At least three factors may cause the annotators to fail this if they are sitting too far from the display, high-resolution display (such as "retina display") is used, or they simply have bad eyesight (and it has not been corrected using glasses etc.). (2) Human annotator can perceive all the pixel they see regardless of the size. (3) There is no loss during the data acquisition in the scanner, meaning that the resolving power of the objective lenses in the scanner should be two times better than the pixel size.

The relationship of the differences between the magnification and resolving power is non-linear. While the image at 400x magnification will appear 4 times better than the one with 100x. The resolving power of 400x magnification is not 4 times better than 100x magnification. In fact, it is only 2.6 times better. Since our approach cannot take magnification into account during the calculation of the DSF, the difference between any two simulated magnification will be quite different from what can be seen in the microscope. This factor should be considered because the FOV usually reflects the amount of image data transmitted to the pathologists, and our digital system cannot achieve the same throughput.

Due to current display hardware limitation, the entire FOV cannot be fitted on the display. In the current OpenHI configuration where the viewer size is set to 1200 by 900 pixel, the viewer can only display 20% and 38% of the tissue surface area compare to what pathologists can see in the microscope. The detail calculation is summarized in 3-4. Noted that this calculation is based on the eye piece lenses that has the field number of 22 mm.

Table 3-4 Comparison of surface area shown in the OpenHI viewer and the microscope.

	OpenHI low	OpenHI high
Area of tissue viewed in the microscope [μm^2]	3,800,000	238,000
Width of the WSI shown in the viewer [μm]	1200×0.728=873.6	1200×0.25=300
Area of tissue viewed in OpenHI viewer [μm^2]	764,000	90,000
% of microscope FOV	20%	38%

To down sample the image, several interpolation techniques may be used including nearest neighbor, linear, cubic, bicubic, etc. In our implementation, we utilize nearest neighbor

interpolation to do the down sampling process since it is the most straightforward method. Although we did not observe any significant changes in appearance of the tissue image once other methods are used, further investigation into this matter may be needed.

3.3 Annotation of ccRCC using OpenHI annotation framework

In this section, we will demonstrate how OpenHI framework can be used to enrich ccRCC diagnostic slides retrieved from TCGA-KIRC project. The goal of the annotation is to acquire high quality annotation from the experts. There are two main topics to be discussed: the framework configuration and guidelines provided to the pathologists or annotators.

3.3.1 OpenHI framework configuration for ccRCC annotation

OpenHI annotation framework should be configured collaboratively between data scientists which will manage the dataset and control the quality of the annotation and pathologists which are is a source of expert knowledge. Parameters will be set including pre-segmentation level and computational cluster—pre-segmented sub-regions which have been segmented using SLIC superpixel segmentation algorithm (c-cluster)—size, grading tier, and viewer size.

1) Pre-segmentation level and c-cluster size

In this work, we arbitrary tried c-cluster sizes ranging from 60 to 20,000 pixel per c-cluster and decided to use three pre-segmentation level with an average c-cluster size of 2,500, 5,000, and 10,000 pixel per c-cluster respectively.

2) Grading tier

According to WHO/ISUP grading system which is a 4-tier system. The grading tier is set to five grades including grade 0 (healthy) and grade 1 through 4 corresponding to the grading system.

3) Viewer size

Setting the viewer size is more complicated than setting other parameters since the size of computer displays are generally limited—the number of pixels need to display full field of view cannot be fitted into the displays. The typical resolution of displays used in this annotation is 1920-by-1080 pixels. We maximize our viewer in the GUI to take up most of the display asset and end up with the viewer size of 1200-by-900 pixels.

3.3.2 Annotation guideline

1) Annotation procedure

There are two stages of annotation: (1) selecting the diagnostic region and (2) annotate cancerous regions. Diagnostic regions are selected once by one annotator to reduce the area

of interest in the WSI. After the diagnostic regions are selected, other annotators will annotate based on the selected regions. If different annotators can select the diagnostic region of their own, it is very unlikely that they will choose similar region since there are a lot of possible and valid diagnostic regions according to the criteria listed in Table 3-5. However, there is one concern about biases from the annotator that choose the diagnostic region and this concern was not addressed in this work since several diagnostic regions should provide the areas diverse enough for machine learning training.

Table 3-5 Criteria of diagnostic regions

Criteria to look for	Criteria to avoid
High cell density (cellular area)	Necrosis
Clear image (well-focused)	Hemorrhage
	Edema
	Fibrosis

Procedures to select the diagnostic region is summarized in Table 3-6. After the diagnostic regions were selected, there should be around 10 to 20 regions of interest. The annotator was instructed to select 10 regions for uncomplicated slides including grade 1, 3, and 4. Twenty regions will be selected if a slide tends to be grade 2 tissue sample to acquire more sample for training.

Table 3-6 Procedure for selecting diagnostic region.

Step	Procedure
1	Go to Magnification Control and select Low.
2	Review the entire slide (under low magnification) to ensure that the slide is of ccRCC1 type to find the diagnostic region.
3	(optional) select High to examine the tissue closely.
4	Click Add current region to record the region. The regions in “To be annotated list” will be updated automatically.
5	Repeat until 10-20 regions are saved.

To perform the second step of annotation, annotators will follow the annotation procedure summarized in Table 3-7.

After the annotation has been made, the annotator has two functions to edit the annotation: (1) undo the latest action using the “Undo” button and (2) remove the annotation by selecting “Healthy” from the Tumor grading menu and right-click on the annotated c-cluster.

Table 3-7 Procedure for selecting diagnostic region.

Step	Procedure
1	Select “Region XX” from “To be annotated list”.
2	Examine the tissue at Low magnification.
3	Go to High magnification to perform the annotation.
4	(optional) Change to bigger PSLV if the c-cluster is too small.
5	Select the appropriate Grade according to WHO/ISUP grading standard from Tumor grading menu.
6	Right-click on the c-cluster to annotate.
7	Repeat from step 5 until the whole diagnostic region is fully annotated.

2) Best practices to resolve conflicting annotation judgment

While our annotation approach where the tissue image is divided into pre-defined regions can improve the efficiency of the annotation, it has also introduced some problems. Those problems will be discussed here. Annotators are instructed to consider nuclei as “cancerous area” in ccRCC grade 1 to 3. In grade 4, cancerous area was not confined to the nuclei since the decision is made based on the tissue pattern.

(1) C-cluster coverage and feature of interest

To cover all cancerous area, all c-clusters that touches—no matter how small—the cancerous area should be selected. This will ensure that the area that has not been annotated belongs to the healthy category. The sample is demonstrated Figure 3-2 where the green area is cancerous area and the yellowed c-clusters should be selected.

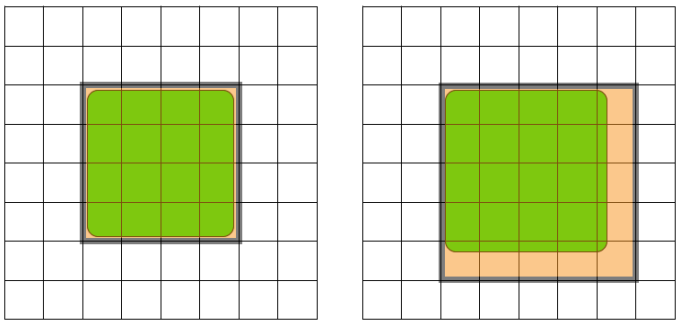


Figure 3-2 Example of c-cluster selection where the green area is cancerous area and yellowed c-cluster are the ones that must be selected.

In grade 1 to 3, annotators should only select the nuclei as shown in Figure 3-3. In some cases, cancerous nuclei may be divided into two or more c-clusters (Figure 3-4). All c-clusters containing the nucleus should be annotated.

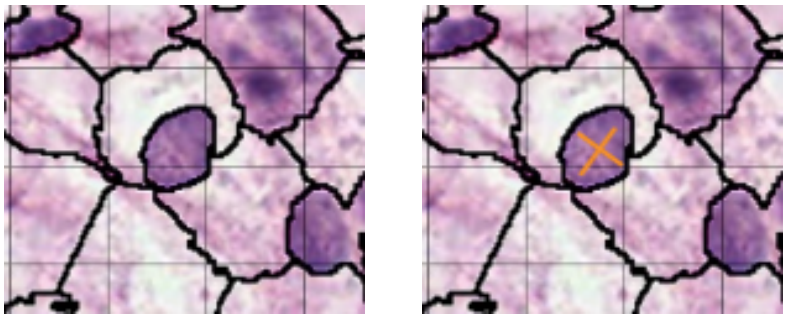


Figure 3-3 Example of a proper nucleus selection when c-cluster fits the nucleus.

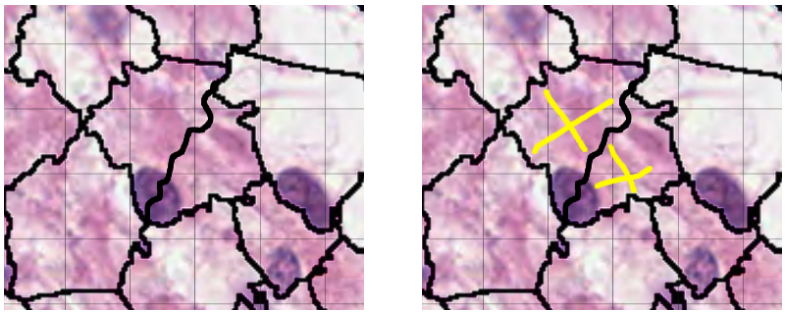


Figure 3-4 Example of a nucleus that was divided into two c-clusters. The yellow marker indicates the c-clusters that must be selected.

(2) Grade selection

In uncomplicated cases, the grade of the c-cluster will be assigned according to the grade of the area inside. In complicated cases, however, the priority of grade selection will be based on the highest grade in each c-cluster. The examples are shown in Figure 3-5 This may cause the system trained based on the annotation to over interpret the tissue grade, but it is better for medical systems to overestimate rather than underestimate the extent of the condition.

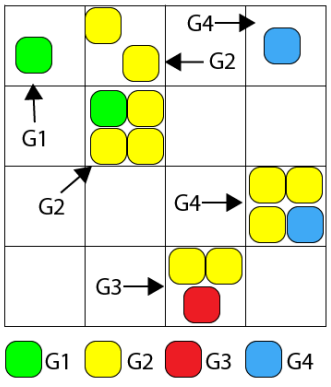


Figure 3-5 Example of a nucleus that was divided into two c-clusters. The yellow marker indicates the c-clusters that must be selected.

(3) PSLV selection

By default, the system will initially set the PSLV to the one with the smallest c-cluster to maximize annotation granularity. However, the c-cluster may be too small for the job and cause

unnecessary effort to select all c-cluster. In that case, the annotator can change to the PSLV with larger c-cluster increase the efficiency of the annotation. In Figure 3-6(a), unnecessary small c-cluster was illustrated. The most appropriate c-cluster size is shown in Figure 3-6(c) which is unlikely to happen. Figure 3-6(b) shows the case where appropriate PSLV was selected. Noted that regardless of PSLV, the annotation should cover similar area and should not affect the ground truth too much. Selecting appropriate PSLV will increase the annotation efficiency and save time. However, there could be some cases where c-cluster is too large for the cancerous area (Figure 3-6(i)) and smaller c-cluster should be selected (Figure 3-6(g, h)).

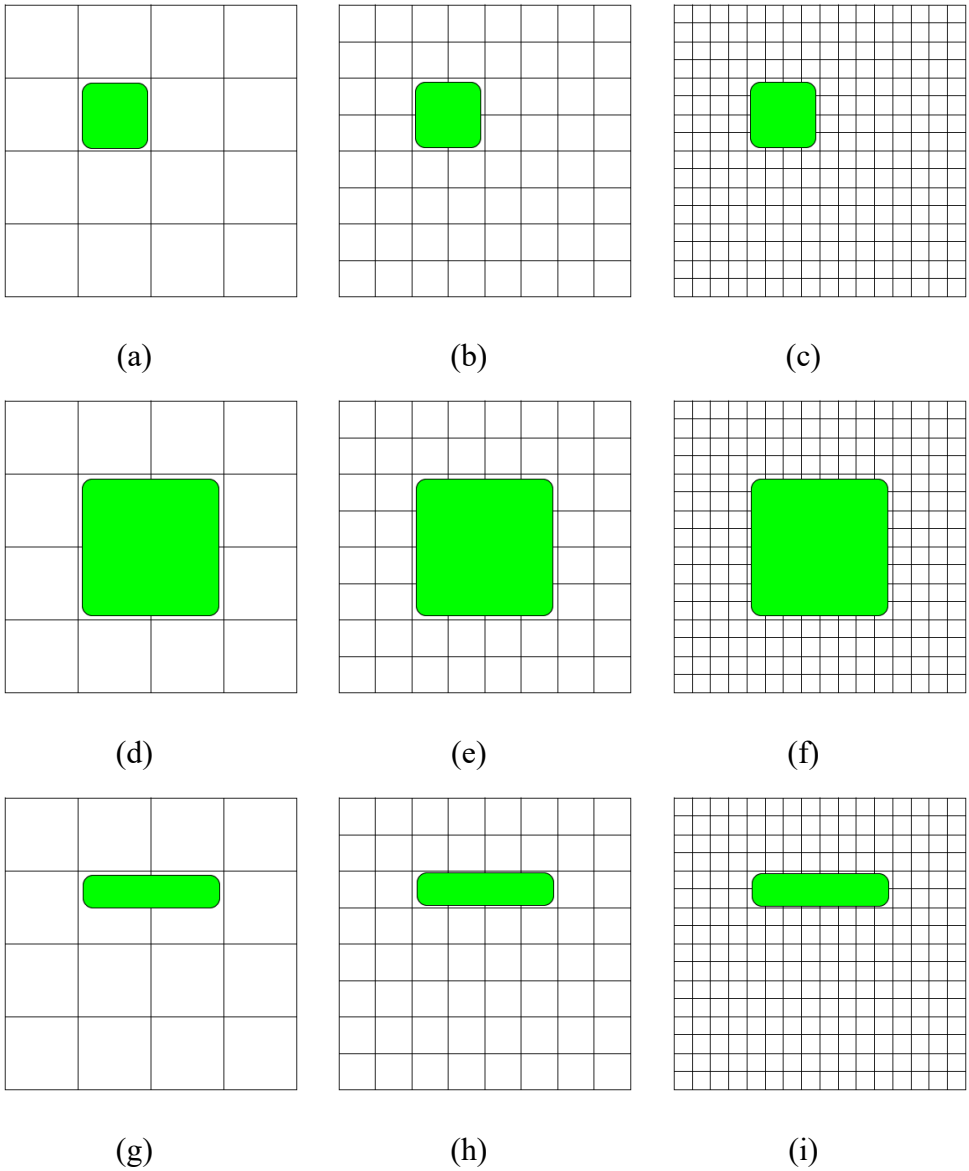
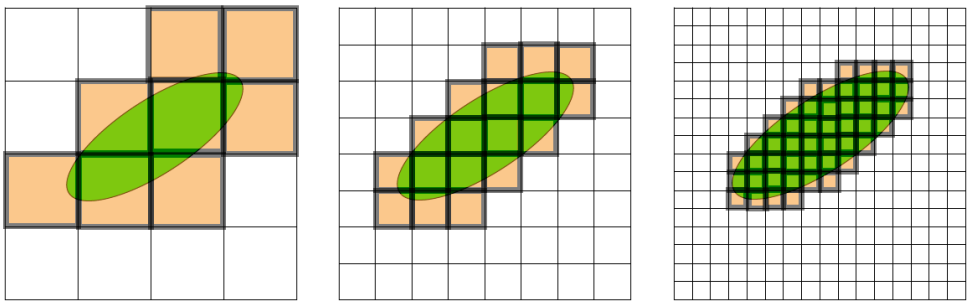


Figure 3-6 Examples of different PSLV setting with varying size of c-cluster: (a-c) are the cases with smaller cancerous area and the c-cluster range from smallest to the biggest PSLV respectively. (d-f) are cases with larger cancerous area. (g-h) are more examples of good PSLV selection while (i) should be avoided. The green region is area of interest.

In the case that c-clusters do not fit the cancerous region, PSLV with the smallest c-cluster should be used to avoid unnecessary annotation of healthy area as demonstrated in Figure 3-7.



(a) (b) (c)

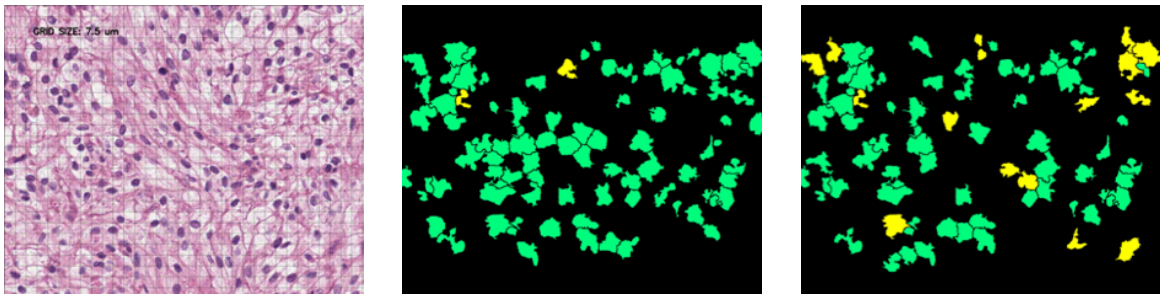
Figure 3-7 The scenario where c-clusters cannot fit the cancerous region and the PSLV with smallest c-cluster (c) should be used and the bigger ones (a-b) should be avoided.

(4) Down sampling factor for low- and high-power digital magnification

As discussed earlier in section 3.2, the digital definition of low- and high-power magnification in the digital system has not been well-defined yet. In this annotation, the down sampling factor for low- and high-power field are set to 2.91 and 1 respectively.

3.4 Sample annotation results

This is an ongoing work; the annotation was set to be made by two pathologists. At the present, annotation on one WSI was complete, resulting in one set of multi-expert annotation. The statistics of the result is the two annotators have made about 20 –50% overlapping annotation, and about 20% of diagnostic regions were annotated. Figure 3-8 shows sample of the annotated results along with original view of the diagnostic region.



(a) Original diagnostic area (b) Annotator 1 (c) Annotator 2

Figure 3-8 Sample annotation result from two annotators.

3.5 Discussion

3.5.1 Usage of annotated dataset

The annotated dataset will be useful in many applications including, but not limited to, nuclei localization, diagnostic area localization, detection of cancerous area—the area with cancerous cells, tissue classification by grade, and to study disagreements between pathologists. In this section, we will the potential usage of the annotations.

1) Nuclei localization

Since low-grade ccRCC rely on nuclear based grading. Each individual nucleus can be distinctly identified as different grade. Our annotations at the cellular level can be used to extract cells and differentiate them from the background: cytoplasm, connective tissues, blood vessels, and other artifacts. The type of annotation that could be labels driving this such system should have markings at the nuclei-level which our annotated dataset has. The annotation of lower grades will benefit this application.

2) Cancerous area detection

The overall detection of cancerous area is also needed to assist pathologists in finding region of interest or diagnostic area. The benefit of such system is to help pathologists thoroughly examine the tissue slide while spending less time on finding the diagnostic regions and more time assessing the grade for the slide. The gist of this kind of system is to assist the pathologists to make better decision, not making precise decision for them.

3) Tissue classification based on tumor grade

Unlike cancerous area detection, this type of system aims to accurately classify the sample grade. The most popular approach is to use deep convolutional neural networks-based model to subsequently classify a small part of the slide. However, it is known that deep neural networks need huge amount of data for training before it can perform the classification well.

3.5.2 Precautions in utilizing any pathologically annotated datasets

1) Disagreements between pathologists

The overall tissue grading is the result of soft decisions. It is common that graders will disagree with each other, even with experienced pathologists. Fueling training-based statistical machine learning systems with multi-expert annotation may result in less than optimal performance as can be seen in (Nir G et al., 2019). Nevertheless, training such system on a single-expert annotation is not ideal either since biases from that particular annotator will have a full impact the prediction outcome. Currently, there are no solutions on how to solve this kind of problem.

2) Patient-level diagnosis needs several tissue slides

As mentioned earlier, the patient-level diagnosis is typically the result of pathologists examining several slides from the same biopsy. The dataset that this work is based on only has one slide per patient. Thus, it may not be appropriate use this dataset for train a system to predict patient-level diagnosis or to study the effect of how several slide-level diagnoses affect the patient-level diagnosis.

4 Whole-slide image analysis for renal cell carcinoma

The current goal of medical image analysis so far is to automate the repetitive tasks which will reduce the workload of specialists and reduce human errors. After the workflow has been automated, it could enhance the diagnostic procedure by unraveling novel features or biomarkers that could be used as an anchor point for disease severity assessment so that appropriate treatment could be administered. More intermediate problem in manual image analysis is about the inter-rater agreement and reliability. A number of investigations [57] have pointed this problem out by performing retrospective studies. A standardized diagnosis results are accomplishable using a single automated system. In short, computers could improve the quality of the diagnosis by increasing the accuracy, sensitivity, and objectivity [4].

To make medical image analysis work, appropriate working environment including the digitization, storage, and processing is needed. For histology slides, the digitization process has been maturing in the past decade with the advance of whole-slide scanners. Before the whole-slide scanner has become widely available, images are taken from the microscope viewing head using a CCD camera. Taking images using the mentioned technique can only acquire a section, as far as the field of view (FOV) of the microscope, of the tissue slide where the area was selected based on the pathologist taking that particular image. Thus, it is technically impossible to acquire the image of the whole tissue slide and get an overview of how the slide may look like. At this technological stage, holistic analysis of the tissue slide is possible and often called whole-slide image (WSI) analysis.

4.1 Problem with medical image analysis on renal cell carcinoma

In the same way as other kind of cancer, renal cell carcinoma can benefit from robust and accurate medical image analysis system. In this study, we specifically study the clear cell renal cell carcinoma (ccRCC). Grading system for tissue sample for this oncological type is strictly magnification-dependent, pathologists were asked to use specific power field in the decision-making process. While the development of digital pathology has introduced digital slides, which is a tissue sample scanned into a digital image and viewed using computer displays, the effect of microscope's magnification has not been proven to be reproducible with digital image processing on computer displays yet. The grading of ccRCC is also controversial (not clear-cut grade description), known to be subjective (but somewhat effective), and requires throughout examination of the slide (not partial examination) [58–61]. Therefore, it is complicate to establish discrete annotation class and analyze them. Furthermore, he current grading standard for ccRCC is created by International Society of Urological Pathology (ISUP) and is not compatible with conventional way of image analysis—extracting a number of comprehensive features

for classification. In short, there are three main problem regarding the current grading system of ccRCC: (1) reliance on microscope's magnification mechanism, (2) subjectivity of current grading system, and (3) lack of underlying biological explanation for each grade. It is urgent that we establish the diagnosis method for ccRCC that works in digital pathology. In this thesis chapter, we will discuss our approach to discover basic parameters or computational visual features that could effectively distinguish the differences among each grade of ccRCC and can directly describe the underlying biological process—such as cell cycle—in cancerous tissue. Our aim is to work with H&E stained image since it is a principle stain for all tissue samples.

4.2 Type of data used in histology image analysis

The current widely used histologic slides for cancer diagnosis is H&E stained slides because of several reasons e.g. being the pathological gold-standard for diagnosis decision, lower cost of preparation, the ability to visualize important pathological features. As a result, there are efforts to put together a competitions such as Camelyon [16], TUPAC [17], ICIAR [18] and public repositories such as TCGA and GTEx regarding this type of data.

4.2.1 H&E staining

H&E stained slides consist of two chemicals and is necessary for all tissues that will be used in cancer diagnosis. Hematoxylin stain binds to nuclei thus make them appear blue or purple. Eosin stain binds to connective tissues surrounding nuclei appear in red, pink, and orange. Although the staining chemicals does not unveil special substances or biochemical interactions. Nevertheless, it provides insights about the morphological changes in the tissue sample. These changes can be both at the cellular, arrangement of the chromatin in the nuclei or overall shape of the cell, or inter-cellular level, structure of how a group of cells arrange themselves. These fundamental characteristics have enabled pathologists to tap into the vast underlying information within the tissue and interpret stages of the disease from biological ques. As well as pathologists, automated analysis framework has also been taking advantage from these visual features.

4.2.2 Immunohistochemistry staining

Immunohistochemistry (IHC) is a type of immunostaining used to highlight specific proteins in the tissue. This is not used in a mainstream clinical routine. Since it is used only in some specific patient cases, small research projects, and clinical trials in precision medicine, large-scale dataset for this type of image has not yet been widely adopted by the machine learning community. Some work has proven that additional information in IHC image can enhance the performance of the automated analysis but further biological explanation of how it can im-

prove the performance beside the fact that it provides additional information is needed. The protein could be selected based on associative studies that indicate correlations between the protein and the oncological sub-type or the stage of the disease. Thus, the way IHC image improves prediction performance can come from the presence, density, or pattern of some stain features and this research question has not been answered at the present.

4.2.3 Tissue microarray

Another popular source of data that could provide specific patterns representing different stage or type of the disease is the tissue microarray (TMA). Compare to WSIs, major difference in the TMAs is the high variety of the patterns since many cases or cores can be fitted to the same slide. Nevertheless, what cores in TMA lack is size. The original purpose of TMA is to perform simultaneous analysis of different molecular targets. Currently, there is one popular TMA repository named The Stanford Tissue Microarray Repository [62]. The information in this repository has enhanced some analysis algorithms [21] to yield better results. Considering the practical situations, TMA is not a standard practice to diagnose the patient. Thus, training the system to recognize specific patterns in the tissue slides cannot rely on this type of information alone.

4.2.4 Use of different data type

In recent years, a number of analysis frameworks have demonstrated that it is possible to construct automated systems to perform pathologists' repetitive tasks with a decent performance [19, 20, 29]. It should be noted that without advance staining, automated analysis can further exploit the underlying information in the tissue slide even with basic staining since the usual routine has time constrain and pathologist cannot fully examine the slide fully. Systems that aim to achieve the mass screening or alleviate pathologists' workload should be able to perform the prediction based on H&E stained whole-slide image.

4.3 Different approaches in tissue slide analysis

Histopathological features can often be quantified into a digitally computable morphological- or texture-based features as decades of conventional machine learning and engineered features has thrived on. Each oncological type has its own characteristic that pathologist could exploit to accurately and conveniently determine the severity of the disease e.g. nuclear-based grading approach is effective and widely used in kidney cancer, mitotic count works well in breast cancer, and determining the number of neurofibrillary tangles in tauopathy. In some sub-type, focusing on the relationship between the cellular activity (disease extent) in cancerous cells and its morphological alterations within the nuclei area should provide insights to engineer

quantitative features that can be utilize the benefit the digital image analysis. Recognition of visual features is crucial to identify oncological type/sub-type and severity of the disease. Changes in visual features are caused by morphological changes in the at the cellular level. This is the result of genotype-tissue expression where various coding DNA strands influence the development and functions of the cell. In cancerous area of the tissue where certain genes are altered, the behavior of the cell would have been changed, so as the morphology of the cell and its nucleus. Some grading standards rely on those cellular patterns. While [63] suggest that the architectural pattern of the cells may also contribute to the assessment. Spotting these features and associating the features with tumor grade are the ultimate goal of automated systems. There are different approaches to perform such tasks. Three major approaches including manual feature engineering, detection and counting, and tissue classification will be discussed in this section along with their advantages and disadvantages. In this thesis, a new approach to analyze the tissue image based on biological comprehension is proposed. Distinct morphological changes at the cellular level can cause the cell to appear differently under the microscope with different staining techniques. Pathologists have exploited these changes so that they could spot unusual activities. Data scientists could use similar approach to craft features that would reflect variation that they wish to spot. The research community has been experimenting with this type of measurement for a long time. As a result, there are thousands of methods to extract features, however the ability to quantify the visual information was not satisfactory. Recent developments in deep learning has unlock this technical bottleneck with the ability of deep learning models to learn complex functions and extract effective features. Nevertheless, the transparency of system is being questioned since intelligence used in the clinic should be explainable.

4.3.1 Analysis based on feature engineering

So far, there have been two approach to come up with good manually crafted features. The first is to directly measure the changes according to biological behavior e.g. measuring the nuclei size and shape using principle component analysis, determining the cell density by measuring the distance between nuclei, and directly determine the amount of stained color in the tissue. In this approach, the analysts need to understand the biological features that pathologists often use to identify the cancerous region in the tissue. The second is to utilize known image processing features to quantify the visual appearance and associate those features with the desired outcome (e.g. grade, severity, and survival outcome). Some researchers [60, 64] have demonstrated this approach and successfully create a system to identify oncological sub-type. In 2016, Yu k. et al. [64] extracted 9,879 quantitative image features from WSIs and TMA slides and successfully use the extracted features to train the regularized machine-learning model and effectively predict the patient's survival outcome. Most of the important

features are texture-based is expected because texture features are highly sensitive to the visual pattern in tissue images. In 2017, an effort to organize the set of image features for 3D radiology scans has been made [28]. A set of quantitative image features has been established and they are grouped into first-order statistics, shape-based (2D and 3D), gray-level cooccurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLZM), neighboring gray tone difference matrix, and gray-level dependence matrix (GLDM) with an average of 15 features per group.

4.3.2 Analysis based on detection and counting

In some oncological type such as breast cancer, the grade of the disease is assigned based on the occurrence of specific entity in the image in a limited area. An example would be the mitotic count in breast biopsy where there are grade 1 to 3 and the count for each grade is less than 7, 8 to 15, and more than 16 at high power fields. In this case, pathologists would have to identify all cells at a mitotic state and count them. The challenge for human interpreter for this kind of task is counting and keeping track of which cell has been counted, while the challenge for machine is to correctly identify the cell that comprise with features that should be recognized as mitotic cells. Another occasion where the detection and counting approach is needed is when special stains are used, and the instance of stained objects must be counted. The automated system that work based on detection and counting should perform better than human by having the ability to keep track of what have been counted and covering more area of the tissue slide, not just some sampled area. It should be noted that the occurrence that were counted has a well-defined boundary or can appear in the tissue independently. Otherwise the counting process would be ambiguous.

4.3.3 Analysis based on tissue classification

Successful application of statistical machine learning, both shallow and deep, is in this area where the analysis pipeline will predict or score each part of the image to reflect the severity of the region. The area of the image fed to the model is arbitrary sectioned. Two popular method to break the WSI down into parts are sliding window and tiling method. In sliding window, arbitrary sub-window size is set along with step size. Sub-windows will be highly overlapped if the step size is relatively small compare to the sub-window size and will have less overlapping area otherwise. This method may generate unnecessary amount of data to interpret. Tiling method is similar to the sliding window, but the step size is equal to the tile size. This sectioning method is somewhat popular and effective, it is used in many recent works [19, 21, 65, 66]. Logically, this approach would be appropriate for applying to grading standard with complex pathological description for each grade e.g. Gleason's pattern and WHO/ISUP grading system since there are pathological terms such as stroma between glands, irregular

masses of neoplastic glands, rhabdoid and sarcomatoid differentiation, nuclear pleomorphism, acinar pattern, eosinophilic hyaline globule pattern, or lymphatic invasion that are too complex to be broken down into a set of interpretable features.

4.3.4 Analysis based on biological comprehension

In this thesis, we propose a novel approach to analyze histopathological image. To promote the transparency of the machine learning system and engineer an explainable system, the machine should analyze the image in the same way that pathologist do. To goal is to engineer an ability for the system to directly quantify the extent of the disease. To do so, the mechanism of how cancer works must be understood. The general characteristics of cancer cells is that the cluster of cancerous cells will have a heightened cellular activity, consuming more energy and nutrients, and create more cells rapidly since the regulation system has been broken. The heightened cellular activity causes the irregular arrangement of euchromatin and heterochromatin in the nucleus which can be spotted using H&E staining and is the histopathological feature that some grading system are built upon. In ISUP grading system for clear cell renal cell carcinoma, the feature is then technically termed “nucleolar prominence” which has high degree of subjectivity.

4.4 Case study: visible nuclear characteristics in H&E stained tissue sample image of low-grade clear cell renal cell carcinoma

In this case study, we hypothesized that the visual effect that WHO/ISUP grading system referred to as “nucleolar prominence” is the combination of nuclei size and intensity contrast within the nuclei surface. After preliminary investigation on the appearance of nuclei from tissue slide with different grade confirmed by pathologists, manually engineered features are then selected and used to determine the extent of the disease for the sample, then the features were used to classify the nuclei grade and the performance was compared to the actual ISUP grading. The WHO/ISUP grading standard for renal cell carcinoma is analyzed according to underlying biological cellular activity.

4.4.1 Materials and methods

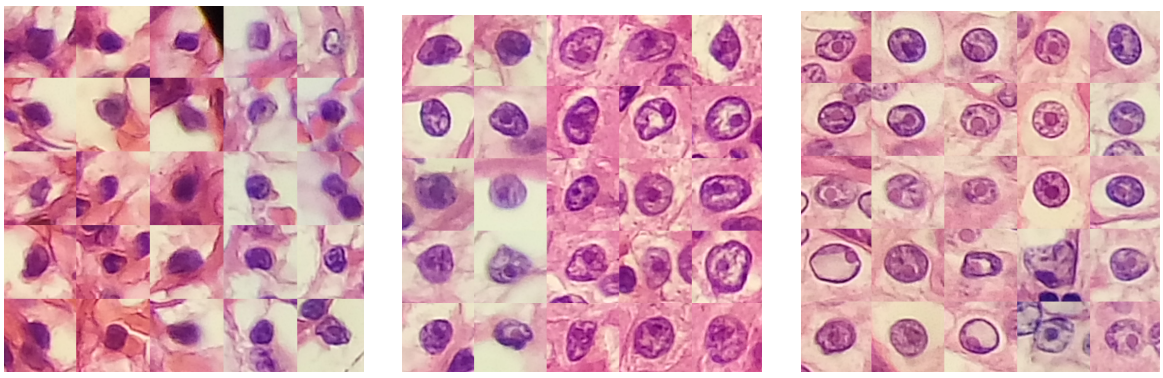
1) Data collection

Digital images of diagnostic slides are acquired via Olympus BX41 light microscope with a smartphone model vivo X6D at the resolution of 8320 by 6240 pixels (50 megapixels) saved in JPEG format from the First Affiliated Hospital of Xi’an Jiaotong University. The image is taken when the microscope is configured to 40x objective lenses and 10x eyepiece lenses (400x magnification). Twenty-three digital images were acquired from 17 tissue slides of 8 patients

that has been diagnosed with ccRCC with pathological grade 1 to 3. The slide-level pathological grade was made by two experienced pathologists. In the case of conflicting grading decision, the highest grade was selected for that specific slide.

2) Motivation and pilot study

Based on our hypothesis that nuclei's size and surface color contrast contribute to nucleolar prominence, we manually extract nuclei from tissue image and arrange them according to the confirmed grade of the tissue slide. Noted that this step is proceed by non-expert. Observation of the extracted nuclei in the pilot study has confirmed our hypothesis that the size and color contrast are the contributing factors to nucleolar prominence. The sample of extracted nuclei according to the slide-level grade is shown in Figure 4-1 and full images can be seen in Appendix B.



(a) ISUP grade 1 (b) ISUP grade 2 (c) ISUP grade 3

Figure 4-1 Sample of extracted nuclei by non-expert according to confirmed slide-level grade from grade 1 (a), 2 (b), and 3 (c). The image of each nuclei has dimension of 91 by 91 pixels.

According to the WHO/ISUP grading standard, there may be some lower-grade nuclei in the tissue slide with higher grade in this set of extracted nuclei since there the slide is graded based on the nuclei with the highest grade. Also, there may be some nuclei with higher grades in lower-grade slide since the slide is very large and pathologists may miss some nuclei. The problem has led us to the design of the study where pathologists will grade each single nucleus individually.

3) Method

One experienced pathologist examined one portion of the image (800-by-600 pixels) at a time and assigned each cancerous nucleus a pathological grade based on ISUP grading system from 1 to 3 using an open source MATLAB script called image-marker (https://gitlab.com/BioAI/image-marker) for the entire field of view (diameter = 550 micron). The marked nuclei are then extracted and grouped based on their grade. Two manually selected features are calculated for each nucleus which are size of surface area as well as the color contrast using root-mean-square (RMS) method (see Equation (4-1)).

$$\sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left(I_{ij}-\bar{I}\right)^2} \tag{4-1}$$

(1) Feature selection

The size of the nuclei is selected as one of the features since we can observe clearly from the sample nuclei that they have different sizes in different grades. Nuclei of grade 1 is significantly smaller than the ones in grade 2 and 3. The size of the nuclei has been studied before in [60] and it has positive relationship with the grade. In this study, we calculate the size of the nuclei by segmenting the nuclei and count the pixels. The second selected feature is color contrast within the nuclei, this is a novel feature. This feature relies on the condensation of heterochromatin—appear blue or purple in the center of nuclei. As cell activity increases, heterochromatin will be condensed into nucleoli and some is pushed to the edge of the nucleus, clearing ways for euchromatin to expand and replicate more DNA strands. The H&E stain does not bind to euchromatin, thus create clear nuclei which is also the origin of the oncological type, "clear cell".

(2) Nuclei segmentation

Since the nuclei are manually localized, segmentation can be made via a set of simple image processing operations. In our segmentation process, we do the following: (1) convert RGB image to grayscale image, (2) binarize the image using global Otsu thresholding, (3) dilate the image (kernel is disk-shaped with r = 5 pixels), (4) flood-fill the image, (5) erode the image using the same kernel as step 3, (6) clear connected components that are connected to the border, (7) select the largest connected component as the segmented nuclei. Regions of interest are then used to calculate nuclei size and color contrast. Example of overall process is shown in Figure 4-2. The segmentation process is implemented in MATLAB. Noted that this segmentation process does not work for all nuclei, thus there are smaller number of segmented nuclei which are 766, 175, and 242 nuclei for grade 1, 2, and 3 respectively.

Figure 4-2 Example of segmentation process including thresholding.

(3) Nuclei classification based on selected features

Two experiments were conducted to train models for nuclei classification to demonstrate capabilities of the two features. The first experiment is based on a multi-class linear support

vector machine (SVM) classifier model in MATLAB version 9.4.0 (R2018a) with Statistics and Machine Learning Toolbox Version 11.3. The second experiment is based on a fully-connected neural networks from MATLAB Neural Network Toolbox version 11.1 where the configuration 2-layer network with 10 nodes in the hidden layer and 3 in the output layer as is illustrated in Figure 4-3.

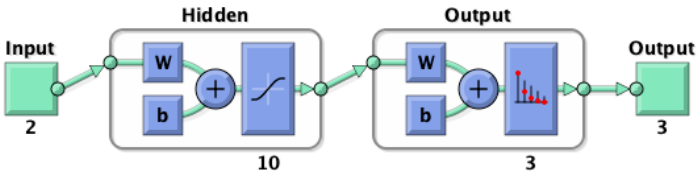


Figure 4-3 Configuration of artificial neural networks for the second experiment.

The input data used and its distribution for training of the two models is plotted in Figure 4-4. The total of 525 nuclei were randomly selected from the segmented nuclei, 175 nuclei for each grade, thus the training data is balanced. In both experiments, the classifiers were trained to predict 3 classes which are nuclei of grade 1, 2, and 3. Distribution of the features from nuclei of different grades are shown in Figure 4-4.

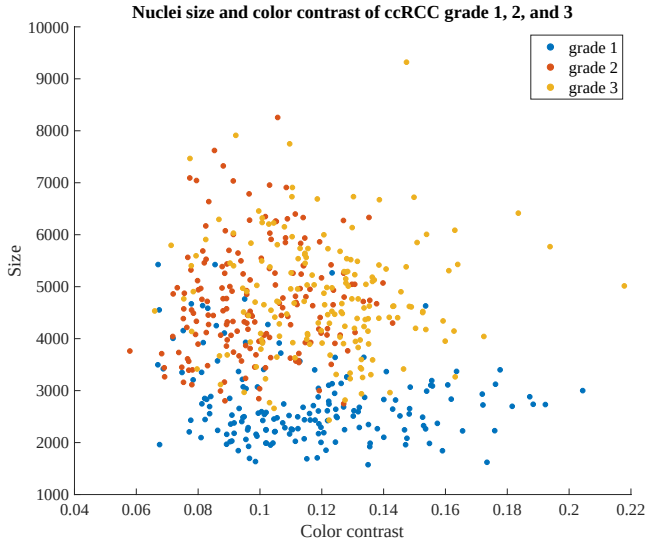


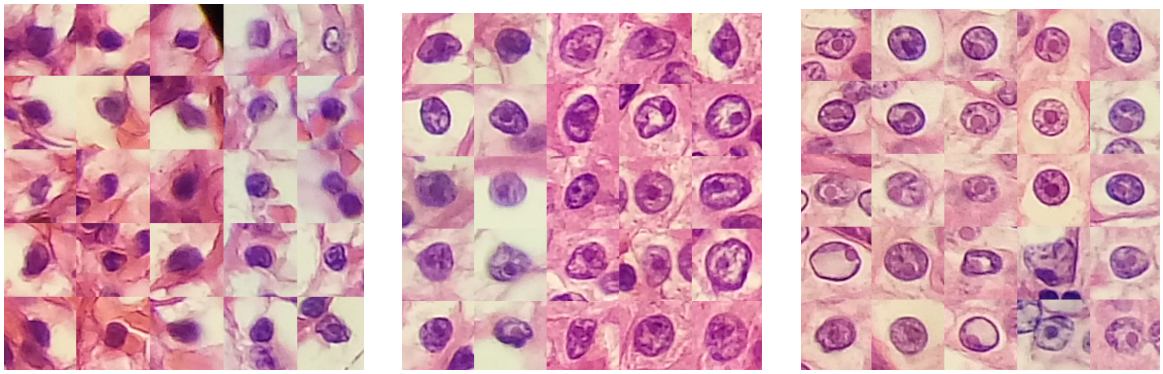
Figure 4-4 Nuclei surface size and intensity contrast of different grade.

4.4.2 Result

1) Annotated nuclei dataset

The dataset is annotated based on the procedure described in the materials and methods section. We have obtained 3,552 nuclei annotated on 66 selected images. In all annotations, 2,619, 723, and 201 nuclei were assigned grade 1, 2, and 3 respectively. It is worth mentioning that in the nuclei in that have been annotated as grade 3, only 15 nuclei came from grade 2

slide and 2 nuclei from grade 1 slide. Examples of graded nuclei are shown in Figure 4-5. It is noticeable that quality of annotation is better than the pilot study. More examples are in Appendix B.



(a) ISUP grade 1 (b) ISUP grade 2 (c) ISUP grade 3

Figure 4-5 Example of nuclei graded by experienced pathologist at nuclei level from grade 1 to 3 (a-c) respectively. The image of each nuclei has dimension of 151 by 151 pixels.

2) Feature analysis

Nuclei size and intensity contrast of the training set are digitally calculated. The statistics of those features are summarized in Table x1. From the digital measurement, the average surface area and normalized RMS value of the nuclei marked with different grade is shown in table x1. It can be seen that nuclei with larger size tend to be assigned higher grades. Similarly, the nuclei with more intensity contrast are likely have higher grade.

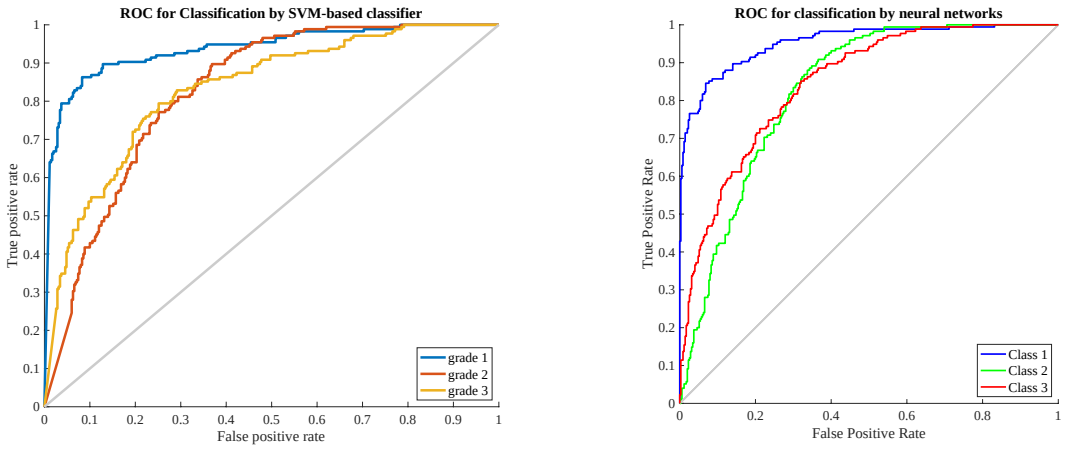
Table 4-1 Averaged nuclei surface area and intensity contrast of the training set

Grade	Averaged nuclei surface area (SD) [pixels]	Averaged nuclei intensity contrast (SD)
1	2744 (753)	0.1164 (0.0292)
2	4710 (1049)	0.0990 (0.0174)
3	4797 (1082)	0.1212 (0.0235)

3) Classification result

The combination of nuclei size and intensity contrast used as digital image features can correctly identify the grade of single nuclei with 71.24% (10-fold cross validation) and 71.80% (overall) accuracy by SVM-based and neural network-based classifiers respectively. Similar performance in the two experiments shows that these classifiers have utilize the full potential of the given features and cannot be further optimized. In the second experiment with neural networks, classifier has better performance in classifying grade 1 nuclei than grade 2 and 3, as the accuracy of classification for different grades are 80.00%, 69.70%, and 65.70% respectively

as shown in confusion matrices in Figure 4-7. Better classification performance on grade 1 versus grade 2 and 3 can also be seen in receiver operating characteristic (ROC) curve for both classifiers as shown in Figure 4-6.



(a) ROC of linear SVM (b) ROC of neural networks

Figure 4-6 Receiver operating characteristic curve for (a) SVM and (b) ANN based classifiers with grade 1, 2, and 3 as a positive class.

4.5 Discussion

4.5.1 Underlying biological process

Based on close examination of nuclei and quantitative analysis, comparing to ultrastructure study using electron microscopy of ccRCC tissue, H&E stained sample examined under light microscope have similar characteristics, only with lower resolution. The fact that euchromatin will get enlarged during heightened cell activity and result in larger nuclei with more densely packed heterochromatin can be seen in electron microscope [67] as well as in H&E stained tissue. In H&E stained tissue, the eosin is stained specifically to the heterochromatin that consist of histone protein. Therefore, the darker area in the H&E image represents heterochromatin and the lighter area represents euchromatin.

4.5.2 Performance of proposed features

Two classifiers chosen to utilize the proposed features have different computational complexity where linear SVM classifier have less complexity and ANN-based classifier has sufficiently complex structure. Yet, they preform similarly in with the proposed features and both have positive results. The evidences should be adequate to conclude that these two features do reflect the nucleolar prominence mentioned in the WHO/ISUP grading standard. The fact that two classifiers have similar performance suggests that features have been used at its full

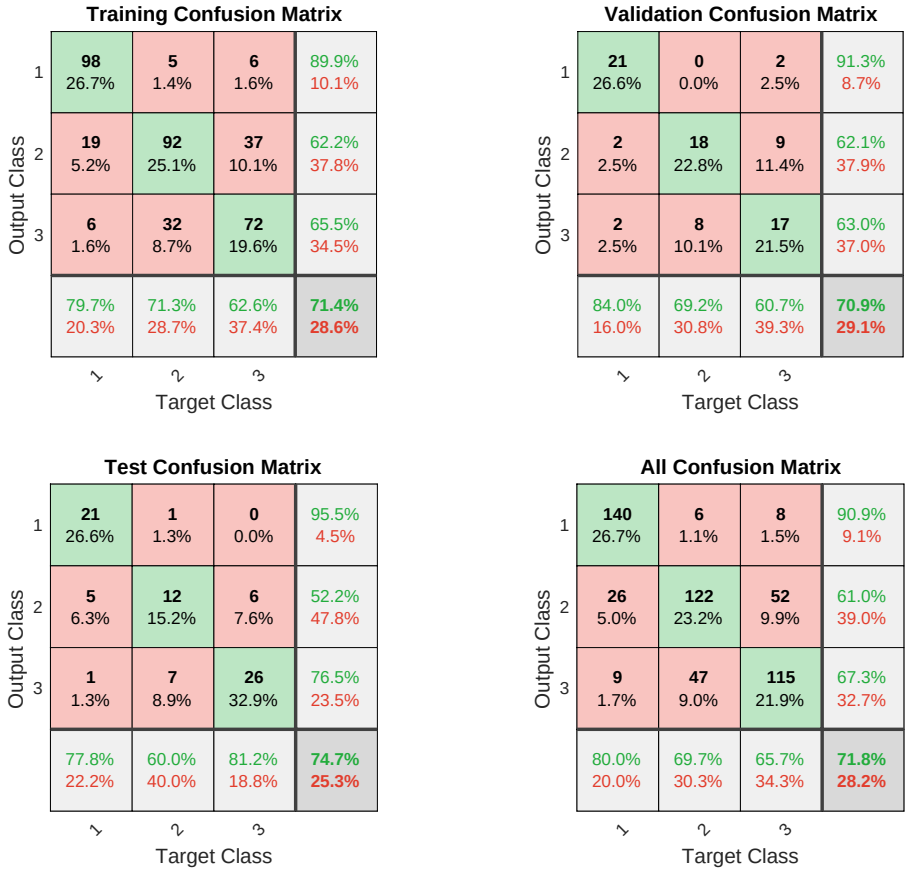


Figure 4-7 Confusion matrices from the artificial neural networks experiment.

potential and the classifiers cannot be further optimized to improve the performance.

4.5.3 Evaluation of the artificial technique on digital pathology

Currently, there are two widely adopted techniques to demonstrate the interoperability and robustness of machine learning models which are unique cross-validation technique and validation on other datasets from completely new sources—hospitals, patient groups, and geographical area. The special property of cross-validation technique used in medicine is often called leave-patient/case-out cross-validation as mentioned by [65]. The gist of this method is to test the variability of data at the patient level since it is known to have high variation. For example, cross-validating the model should completely hold all data from one or more individual patient out during the training period and test using the held-out data. Using different datasets may also be effective in validating machine learning models. [68] have demonstrated associating multi-omics features from two datasets including TCGA cohort and Mayo Clinic cohort.

4.5.4 Usage of prediction results and analysis results

Prediction results from automated systems can be utilized in many ways and there are pros and cons in each approach. While the ultimate goal of automatically grading histopathological slides is to achieve patient-level diagnosis which could be done by fully automated systems, other usage of predictions is also useful. For instance, computer-assisted diagnosis could shave some repetitive and simple tasks off, and let the pathologists focuses on harder part of the diagnosis. A fully automated system to predict patient-level diagnosis is ideal. However, it is not considered safe to utilize such system since any number of preventable false negatives is not acceptable. Moreover, a patient-level diagnosis could be considered as a complex decision since it involves multiple source of data—parts of slides, this makes the demand for even larger amount of data and researchers have agreed that current magnitude of data is not enough. For example, number of cases in top 6 sub-projects in TCGA are around 500 cases. This number is far from enough since there could be thousands of new cancer cases just in the US every day. Compiling medical records from every day's diagnosis in hospitals would requires standardization and it is not feasible to do so at the moment due to lack of international standards and biomedical infrastructure. More imminent use of automated systems falls in the area where low-level decisions are made. These are spotting cancerous segments of a slide—without predicting grades, localize and count certain indicative elements—reduce repetitive tasks, and find similar area of interest. However, CAD should be evaluated if the systems cause biases in the diagnosis.

4.5.5 Risks in utilizing machine learned information in the pathological diagnosis

All automated systems are somehow derived from a training, development, and testing datasets. Those could be different by randomization, held-out methods, or originating sources. A system designed for certain type of disease may assumed that it would work with data from any settings. This assumption should be ceased until it is proven that data used has similar or same characteristics. For example, tissue stained by the same staining technique may appear differently. Immediate response to tackle this problem is to release datasets used in the development process, thus users could have ideas if their input data is compatible with the system. A more long-term solution may be to develop tissue characteristic similarity tests, and they should be able to spot the difference between patients, data cohorts, hospitals, and geographical areas. Performance matrices for this kind of automated systems should be evaluated as well. Conventional methods to report machine learning models such as accuracy, sensitivity, specificity, precision, recall, etc., may not be enough to evaluate medical systems. The use of patient survival analysis: disease-free survival and overall survival, using hazard ratio or

Kaplan–Meier estimator may be more appropriate. Finally, users of automated systems should understand these matrices and know how they are developed to avoid pitfalls that would make their use of artificial intelligent system fail and cause erroneous diagnosis.

5 Conclusions and suggestions

In this thesis, we proposed three main topics including an introduction to histopathological image annotation platform—OpenHI, a discussion on best practices for annotating clear cell renal cell carcinoma, and a new approach to analyze tissue slides based on underlying biological interpretation with preliminary demonstration. This is done in response to the need for semantic understanding of histopathological pattern which will drive the development of artificial intelligence techniques in digital pathology. In addition, this chapter also provides a solution on how to keep up with cancer research which is a fast-growing research area. Digital pathology is a relatively new area of study within cancer research. We believe that addition of our annotation software and creation of large-scale annotated dataset may induce many new ideas and lines of work, these are discussed in the outlook section.

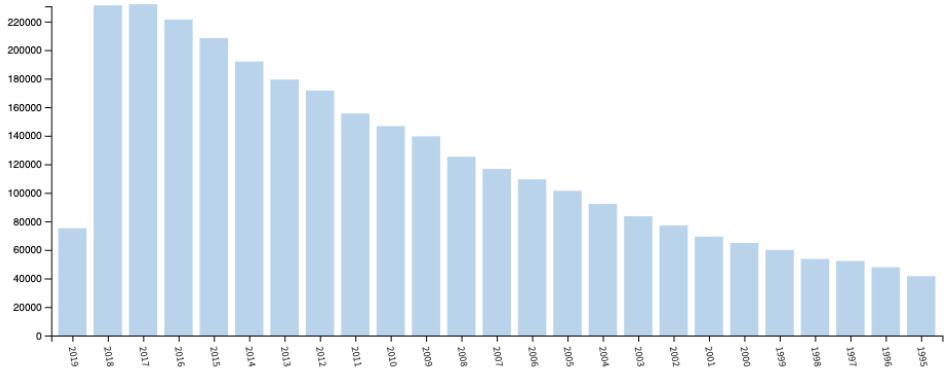
5.1 Rapid development in cancer research

Cancer research is a fast growing and has become multidisciplinary research area. It is challenging to keep up with all literatures published recently and sort things out. Figure 5-1 shows the number of publications according to Clarivate Analytics in the past 25 years based on three keywords: cancer, whole-slide image, and histopathological image.

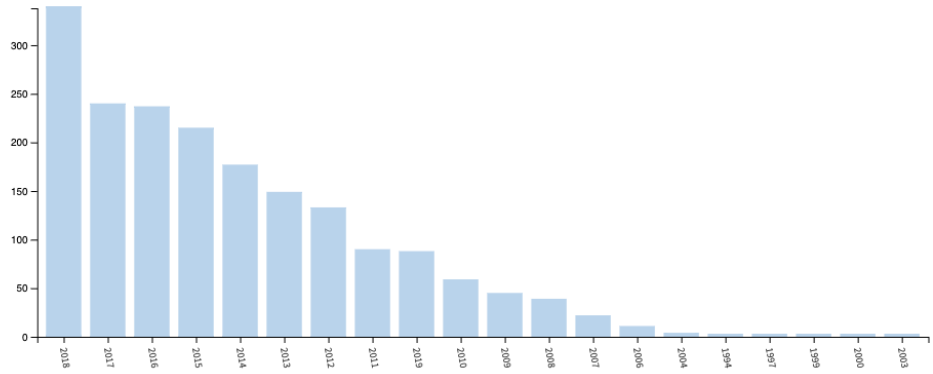
It is crucial to study and follow international/national organizations and their standards. The example of main organizations for cancer in general and for some particular oncological type including renal and lung cancer in Table 5-1. Noted that there are no main and universal organization responsible for all types of cancer nor all regions in the world. Intelligent systems that aims to assist cancer screening and diagnosis should comply with the existing standards and flexible enough to adjust to new guidelines.

5.2 Histopathological image platform

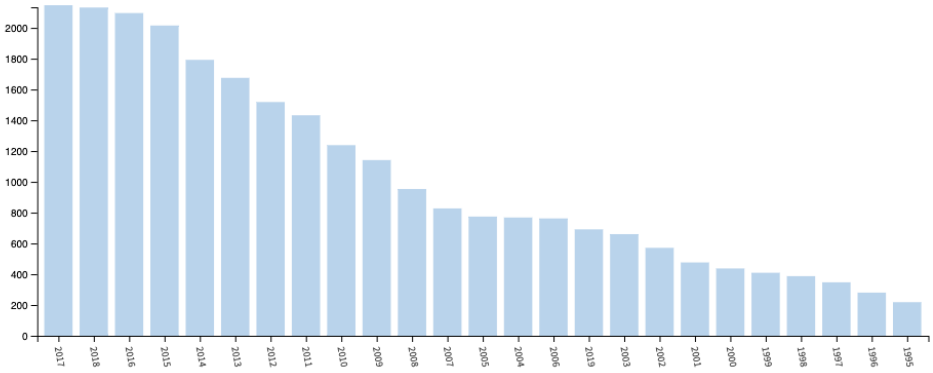
Digitalized histopathological images are increasing in a fast pace with continuous health informatics development around the world. The images present phenotypes of tumors at cellular level and may support the association study with genotypes from sequence data. OpenHI may accelerate precise creation of phenotype annotations with semantic meaning in the images. Additionally, the framework utilizes web technology, therefore is capable of collaborative annotation which is a foundation of crowed-sourcing to create large-dataset. As a result, large-scale datasets with precise and semantically rich annotations which is suitable for training computational model could be efficiently created. The framework is open source and could be easily extended and implemented into a clinical decision-making workflow [48]. It also can be



(a) keyword: cancer



(b) keyword: whole-slide image



(c) keyword: histopathological image

Figure 5-1 Trend of publications in the past 25 years based on three keywords: (a) cancer, (b) whole-slide image, (c) histopathological image.

easily configurable at the back-end for the data scientist to adapt different diagnosis standards, e.g., various cancer sub-types or gradings.

Table 5-1 International organizations responsible for regulating medical procedures and guidelines related to cancer. Noted that this is not an exhaustive list.

Name [Abbreviation]	Disease area	Target	Responsibility	Resources
World Health Organization [WHO]	General	General population	Identify and maintain a list of carcinogenic substances and provide guidelines for the general population. WHO cancer report International Agency for Research on Cancer [IARC] General General population Work as part of WHO to collect and publish global statistics on cancer situation.	WHO “Blue Book”: overview guideline to various cancer types
American Joint Committee on Cancer [AJCC]	General	Clinicians	Maintain cancer staging guideline.	Cancer staging guideline
US National Institute of Health and National Cancer Institute [US NIH & NCI]	General	US population and re-searchers	Provide funding for research activity promoting well-being in the US.	TCGA Project, GTEx Project, Genomic data commons (GDC) repository, etc.
International Society of Urological Pathology [ISUP]	Urology	Pathologists	Maintain cancer grading guideline for urinary system.	ISUP grading guideline for kidney and prostate cancer
The United States & Canadian Academy of Pathology [USCAP]	General	Experts	Maintain guidelines related to pathology	USCAP pathology report for renal cell carcinoma, etc.
International Association for the Study of Lung Cancer [IASLC]	Thoracic	Experts and patients	Enhance the understanding of lung cancer among scientists, members of the medical community and the public	Classification guideline for lung adenocarcinoma, etc.

This work does not only propose a framework for annotating WSI which overcomes technological challenges to read and store extremely large histopathological image, but also discuss procedures to combine semantic meaning within existing cancer grading systems onto digital slides. The proposed process along with the framework complies with usual pathological rou-

tine thus pathologists can effectively, efficiently annotate using the framework to achieve high quality biomedical datasets.

Large-scale datasets with precise annotations may be efficiently created by the framework. Artificial intelligent methods, for example, based on statistical machine learning, could benefit from the rich features in the data and move forward to practically assist the pathologist's routine laboratory work. Such pipeline could also provide a solution to imminent issue such as misgrading which could lead to misdiagnosis and to provide a good foundation for the future development of phenotype-genotype or multi-omics associations [30].

5.2.1 Current problem on inter-rater reliability test

In renal cancer, the grading system is based on nucleolar prominence which is a highly subjective and controversial criteria. Nevertheless, the prominence of the nucleoli has proven itself to be useful and be able to distinguish patients of different grades [60]. Discordance between pathologists is not a problem specific to raters in the virtual slide environment [69]. The problem also exists in the setting of conventional glass slides as can be seen in the study for breast cancer [57]. Introducing more than one annotator to establish multi-expert annotation does not entirely solve the disagreements between pathologists since there are no best way to combine the multi-expert annotation. This problem can be divided into two parts: defining the spatial boundary of regions from multiple sources and deciding the final grade for the same region. It can be claimed that no two annotators can select the exact same region down to the pixel-level based on freehand region selection and often the degree of differences is large. Negotiating the middle ground between multiple annotations is hectic. Selecting the maximum coverage is good for minimizing false negatives but increase the chance of false positives. Similarly, in deciding the final grade, maximum grade can be selected and increase a risk to produce false positive results. In different situations, less experienced raters may under or overestimate the grade of the tissue [45, 57]. It is proven that machine has a better time understanding single-expert annotation [65]. An automated system with good performance trained with single-expert annotation might be the result of non-generalized system.

5.2.2 Effect of image digitization and digital visualization on annotator's judgements

As mentioned earlier, in Section 2.5.2, the digital representation of WSI is far from perfect and not equivalent to an image that pathologists would experience via conventional microscopes. Magnification-dependent grading systems may suffer further irreproducibility from this matter. The effect of resolving power on the judgement of pathologists needs further study. The problem regarding this matter could be broken down into three aspects including digital

image production (scanning), storing file format, and visualization.

1）Glass slide image acquisition

Whole-slide images are scanned by whole-slide scanner. Only some part of information stating the components used during the scan was recorded to the digital file as metadata. The crucial piece of information used in this project is the pixel size. This parameter state how large is the pixel in the WSI on the physical glass slide. However, this important parameter is the result of magnifications from objective lenses and scanner sensor size. Since whole-slide scanner is a closed system, further analysis on the effect of objective lenses' resolving power and the scanning sensor size cannot be studied and users of WSIs must blindly trust scanner's manufactures that they have deliver the pixel value according to the real physical color on the glass slide without any compromises in quality. The effect of resolving power on image digitization should be closely studied and the parameter pixel size should work in the same way as Rayleigh resolution limit.

2) Whole-slide image file storage

Usually digital files do not deteriorate. But the loss in quality occur when the image is first saved from raw files to compressed format. Since WSIs generally have large file size, it is necessary to save WSI files in economical ways. JPEG-2000 compression algorithm is usually used to compress histological slides since they are considered to be natural image (not computer graphics where PNG format is more suitable). There are degrees of compression in both bit-depth and resolution. So far, there are no evidences that WSIs are compressed in term of resolution, but only bit-depth. One study [70] has investigated in this matter and find out that the compression does not affect the interpretation of deep learning algorithms. Nevertheless, further study on the effect of JPEG compression should be made, finding optimized point of compression to store the file [71]. Alternative lossless compression algorithm should also be explored as [72] suggested.

3) Whole-slide image visualization

Recently, US-FDA has published a regulatory guideline for whole-slide imaging system [51]. The guideline emphasizes on delivering physical color of glass slides to the monitor, the idea was named imaging pathway. However, the system does not take how the user of the system will perceive the image into consideration. This effect is hard to study since the effect may vary from person to person. According to [50], the size of computer display has a lot of effect on the magnification of the image itself. Human eye's resolution and field of vision should also be taken into consideration.

5.2.3 Performance of the software

Regarding the performance of the annotation framework, most of the processing time is spent on retrieving and compiling different layers of images to visualize the current state of a particular part of the WSI. The pre-loading operations also take long time since the entire boundary matrix must be loaded into the memory, larger WSI will cause longer loading time, thus more framework initialization time and more memory needed. Part of the reason causing this bottleneck is that conventional image file format and read/write operations were used to accomplish these tasks and they were not designed for images with such large dimensions. The process is illustrated in Figure 5-2.

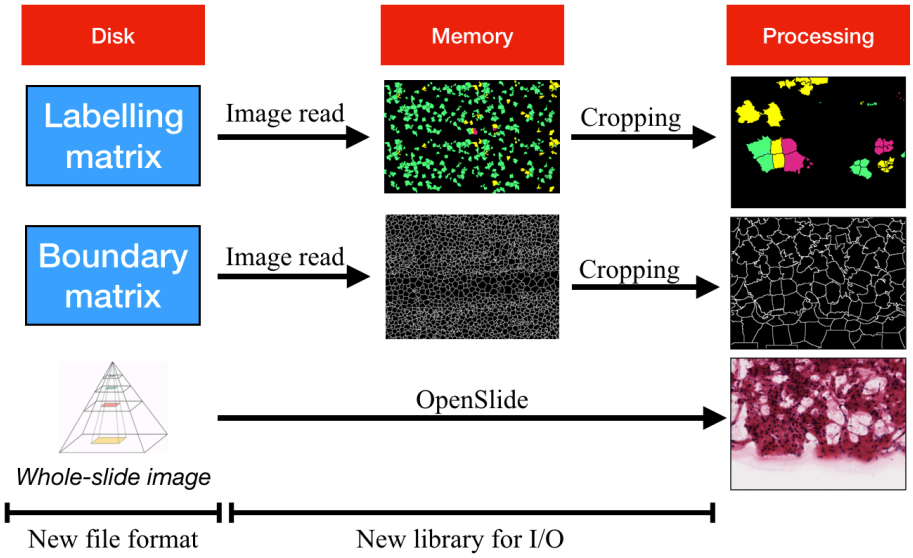


Figure 5-2 Schema of whole-slide image file reading and visualization.

Tackling this problem, we suggest that a new file format with support to store annotation data is needed. This could be a completely new file format based on multi-scale image techniques or deep zooming technology [73]. It could also be an implementation of long proposed, but not popularized, DICOM standard for light microscopy, specifically for WSIs [74]. Together with the file format, tools with programmatic interface similar to OpenSlide should be provided, ensuring that the tool can be used on all computing platforms. The two new components of the system would replace complicate low-level operations for prospective users. As in Figure 5-2, it could be noticed that this is one-way operation, writing back to original WSI files is not possible at the moment. It is suggested that new systems should support writing and modification as well.

5.3 Annotated dataset

The proposed annotation framework was used to annotate WSIs from TCGA-KIRC project. The annotation was made according to WHO/ISUP grading standard at the c-cluster level—adjustable by annotator. Digital tissue slides are shown to annotators on computer displays with certain resolutions for low- and high-power field. The appropriate down-sampled version of original WSIs are processed by our method based on microscopes' resolving power, Rayleigh resolution limit, and Nyquist sampling theorem. Annotations were made by two experienced pathologists.

The developing dataset with annotations provided by experts would not only be able to drive the development of statistical machine learning algorithms, but also be a gold-standard for more annotations for histopathological slides to be made. Annotations can also be used for pathologist training purposes, providing interactive training where students could trace how professional pathologists interact with slides and discover important features. The tracing is possible by looking into log files and reproduce annotation events.

Digital tissue slides with granular annotations of digital tissue slides could be utilized to solve many problems including, but not limited to, cancerous nuclei and region localization, nuclei or cell segmentation, tissue grade classification, and patients' cancer diagnosis prediction. Diagnosis discordance is a known problem among pathologists grading tissue slides, but the disagreement occurs at the patient- or slide-level. Granular digital annotation may be able to provide insights to where were disagreements originated.

5.4 Whole-slide image analysis for renal cell carcinoma

The analysis of renal cell carcinoma can be established in many ways depends on type of input data and analysis approach. Input data in the form of WSI could be H&E stained slides, IHC slides, or a bunch of TMAs. Each of these has its own characteristics, and the most popular one is H&E slides which is considered to be a principle stain and all tissue samples must go through this staining procedure. It is easier to acquire large-scale data of the staining type and mass analysis would make a better impact to the society. Different analysis approaches have been tested out in the past years. They are based on feature engineering, detection and counting, and tissue classification. The latter two could benefit from a rapid development of deep learning algorithms. Nevertheless, in this thesis, we demonstrate a relative new approach based on biological comprehension; examining the underlying cellular phenomenon based on existing pathological grading systems then find appropriate visual features to quantize the visual information according to the extent of the disease. With only two manually selected features based on underlying cellular biological process, a simple statistical machine learning models can achieve the classification accuracy of 70%. At the end, we discuss about the advantages of

the proposed approach to analyze histopathological images. There are also discussions about usage of machine learning techniques in clinical settings and risks involved and how to avoid them.

5.5 Outlook

5.5.1 Visible cellular patterns semantic networks

We believe that cellular patterns seen by pathologists do have meanings. They could be translated into different forms: tumor grades, certain pathological terms, indicators for biological processes, etc.

5.5.2 Association of histopathological patterns with patient's outcome and oncological sub-types.

Although biological explanation on relationship between cancer causing genes and visible visual patterns and identification in histopathological analysis. The conventional method for establishing grading systems involve retrospective study with a sizable study cohort from different hospitals focusing on patient survival analysis, often by Kaplan–Meier estimator and hazard ratio. Grading systems are settled by grouping pathologically identified patterns and associate them with severity level, thus creating a tiering systems, this process could be seen clearly in [63]. Visual cellular patterns may be extracted with a more complex and computable techniques such as deep learning, and it may yield better results. Similar approach can be seen in [68] with non-small cell lung adenocarcinoma. This could be done instead of training classifiers against conventional grading systems, but the effects on diagnosis should be further studied.

Similar association has been made by [20, 64], associating sub-types of non-small cell lung cancer (NSCLC) which could only be identified by genetic tests with histopathological patterns. Both experiments show positive result and proof the feasibility of the fact that there are hidden patterns that are indicative of cancer sub-types and they have not been exploited in conventional pathological analysis yet. Besides discovering new visual features, cellular histopathological features may be established into computable or quantizable amount by gathering well-established image features summarized in Table 5 2 by feature category in for radiology scans [28]. Similar approach could be organized to work with histopathological images.

Table 5-2 Samples of radiomic features that works with 2-dimensional images categorized by image processing techniques [28]

Category	Sample features	Number of total features
First order statistics	Entropy	19
	Minimum	
	ith percentile	
	Mean	
	Median	
	Root mean squared	
	Standard deviation	
Shape-based	Perimeter	10
	Sphericity	
	Major Axis Length	
	Elongation	
Gray level cooccurrence matrix	Correlation	24
	Difference average	
	Contrast	
	Difference entropy	
Gray level run length matrix	Gray Level Non-Uniformity (GLN)	16
	Long Run Emphasis (LRE)	
Gray level size zone matrix	Zone Percentage (ZP)	16
	Gray Level Variance (GLV)	
	Large Area High Gray Level Emphasis (LAHGLE)	
Gray level dependence matrix	Large/Small Dependence Emphasis (LDE/SDE)	14
	High Gray Level Emphasis (HGLE)	
	Dependence Variance (DV)	

5.5.3 Association of digital slides with other data modalities

Three categories of data have always been biologically connected: genotype, phenotype, and clinical records. They have influence on one another. Digital slides are part of phenotype, derived from tissue samples extracted from patients by biopsies. They could be seen by pathol-

ogists under light microscope or digitally recorded as WSIs. Different tissue slide preparation would yield different appearance such as different tissue cutting and staining techniques. After all, the emerged visible patterns are the result of gene—normal or cancerous—expressions. The expressed patterns would also have influenced how a group of cells or an organ works, which could be observed externally or experienced by the patients as symptoms. A more crude but effective method to examine patient's internal organs is radiology scans where tumors could be detected, yet it is caused by unusual growth of cells accumulated into tissue. Clearly, they are all connected as shown in Figure 5-3.

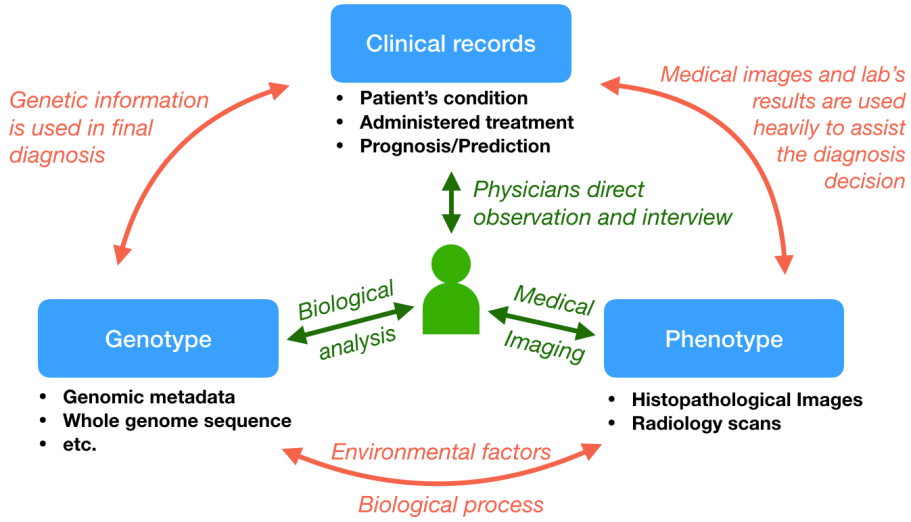


Figure 5-3 Illustration of connections within different data modalities in medicine.

Biologists would have a relatively crude understanding of how this mechanism works in detail compare to what could be achieved by machine. This is the case because human can only perceive limited amount of information at a given time, leading to either close examination of one particular part of the data and ignoring holistic view or vice versa. Examples of data that human cannot perceive are raw DNA/RNA sequence, subtle patterns in histology slides, and clinical records over a long time. Computer systems, on the other hand, can perceive and analyze an entire set of data, the remaining problem is what do we—the user of the systems—want to do—extract, interpret, analyze, etc.—with the data.

Acknowledgements

First of all, I would first like to thank my inspirational advisor Professor Chen Li for supervising the project and providing continuous support at every turn. He has had created an enjoyable research environment for all team members.

I would like to thank scholarship three important organizing committees at Xi'an Jiaotong University, the Consulate General of Royal Thai in Xi'an, and the National Science and Technology Development Agency (NSTDA) in Thailand which offer the Xi'an Jiaotong University Princess Sirindhon of Thailand Special Scholarship in conjunction with Chinese Government Scholarship Council. This scholarship has provided all the support I needed to complete the study for this degree. It is privileged to be a part of student program in the Information Technology Foundation under the Initiative of Her Royal Highness Princess Maha Chakri Sirindhon with a mission to broaden and utilize knowledge in the field of information technology to benefit the development of humanity. I am thankful for all additional help and support from the Anandamahidol Foundation. In the city of Xi'an, the team at Royal Thai Consulate-General has been very kind and supportive. They have made adjustment to life in China much easier.

For the record, this work has been supported by The National Key Research and Development Program of China (No. 2018YFC0910404); National Natural Science Foundation of China (Grant NO: 61772409); Ministry of Education-Research Foundation of China Mobile Communication Corp (MCM20160404); The consulting research project of Chinese Academy of Engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China"; Project of China Knowledge Centre for Engineering Science and Technology; Innovation team of Ministry of Education (IRT17R86); Innovative Research Group of the Nation Natural Science Foundation of China (61721002).

I also would like to thank professors and friends from various external organizations including the Department of Biomedical Engineering and the Faculty of Medicine at Srinakharinwirot University, the Faculty of Medicine at Chiang Mai University, Guangdong Lung Cancer Institute and Chinese Thoracic Oncology Group in Guangzhou, and Mayo Clinic. They have been supportive. A lot of valuable academic advice were kindly given, steering this research project into the right direction.

Regarding people in Xi'an, I would like to thank the team members in the Biomedical Semantic Understanding Group for their help, support, and friendship. A close watch from Doctor Wang Chunbao and Professor Guanjun Zhang in the Depart of Pathology, the First Affiliated Hospital of Xi'an Jiaotong University was highly appreciated. Although I have spent just a short while in the quantum optics laboratory, I must also mention all the support from friends in the laboratory of Professor Zhang Yanpeng who help me to better navigate Chinese

Culture and Xi'an City in the early days of my master study. I am grateful to all my international and Chinese friends outside the lab. In addition, I think Thai senior and junior students for friendship.

Special thank-you note to those who made the LaTeX thesis template available for Xi'an Jiaotong University. It has made life much easier in formatting this thesis, saving me from undocumented template.

Finally, I must express my profound gratitude to my parents for providing me with unconditional love and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Besides, I wish to express my appreciation to my girlfriend who has been supportive throughout the years.

Xi'an, China
May 2019
Pargorn Puttapirat

References

[1] Sharma G, Carter A. Artificial intelligence and the pathologist: Future frenemies?[J]. Archives of Pathology and Laboratory Medicine. 2017, 141(5):622–623. DOI: 10.5858/arpa.2016-0593-ED.

[2] Siegel R, Miller K D, Ahmedin J. Cancer Statistics[J]. Ca Cancer Journal. 2017, 67(1):7–30. DOI: 10.3322/caac.21387.

[3] Allard B, Aspeslagh S, Garaud S, et al. Immuno-oncology-101: Overview of major concepts and translational perspectives[J/OL]. Seminars in Cancer Biology. 2018, 52(February):1–11. https://doi.org/10.1016/j.semcancer.2018.02.005. DOI: 10.1016/j.semcancer.2018.02.005.

[4] Meijering E, Carpenter A E, Peng H, et al. Imagining the future of bioimage analysis[J]. Nature Biotechnology. 2016, 34(12):1250–1255. DOI: 10.1038/nbt.3722.

[5] Evans A J, Bauer T W, Bui M M, et al. US Food and Drug Administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised[J]. Archives of Pathology and Laboratory Medicine. 2018, 142(11):1383–1387. DOI: 10.5858/arpa.2017-0496-CP.

[6] Gilbertson J R, Ho J, Anthony L, et al. Primary histologic diagnosis using automated whole slide imaging: a validation study.[J/OL]. BMC clinical pathology. 2006, 6: 4. http://www.ncbi.nlm.nih.gov/pubmed/16643664http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1525169. DOI: 10.1186/1472-6890-6-4.

[7] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.[J/OL]. Contemporary oncology (Poznan, Poland). 2015, 19(1A): A68–77. http://www.ncbi.nlm.nih.gov/pubmed/25691825http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4322527. DOI: 10.5114/wo.2014.47136.

[8] Keen J C, Moore H M. The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine.[J/OL]. Journal of personalized medicine. 2015, 5(1):22–9. http://www.ncbi.nlm.nih.gov/pubmed/25809799http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4384056. DOI: 10.3390/jpm5010022.

[9] Goode A, Gilbert B, Harkes J, et al. OpenSlide: A vendor-neutral software foundation for digital pathology[J/OL]. Journal of Pathology Informatics. 2013, 4(1):27. http://www.jpathinformatics.org/text.asp?2013/4/1/27/119005. DOI: 10.4103/2153-3539.119005.

[10] Litjens G. Automated Slide Analysis Platform (ASAP) [M], 2015.

[11] Bankhead P, Loughrey M B, Fernández J A, et al. QuPath: Open source software for digital pathology image analysis[J]. Scientific Reports. 2017. DOI: 10.1038/s41598-017-17204-5.

[12] Gutman D A, Khalilia M, Lee S, et al. The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research[J]. Cancer Research. 2017, 77(21):e75–e78. DOI: 10.1158/0008-5472.CAN-17-0629.

[13] OpenSeadragon Project [M], 2013.

[14] Schneider C A, Rasband W S, Eliceiri K W. NIH Image to ImageJ: 25 years of image analysis[J/OL]. Nature Methods. 2012, 9(7):671–675. http://www.nature.com/articles/nmeth.2089. DOI: 10.1038/nmeth.2089.

[15] Della Mea V, Baroni G L, Pilutti D, et al. SlideJ: An ImageJ plugin for automated processing of whole slide images[J]. PLoS ONE. 2017, 12(7):1–9. DOI: 10.1371/journal.pone.0180540.

[16] Bejnordi B E, Veta M, Van Diest P J, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer[J]. JAMA - Journal of the American Medical Association. 2017, 318(22):2199–2210. DOI: 10.1001/jama.2017.14585.

[17] Veta M, Heng Y J, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge[M/OL]. 2018: 1–22. https://arxiv.org/abs/1807.08284.

[18] Aresta G, Araújo T, Kwok S, et al. BACH: Grand Challenge on Breast Cancer Histology Images[M/OL]. 2018. http://arxiv.org/abs/1808.04277.

[19] Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent[J/OL]. Scientific Reports. 2017, 7(March):1–14. http://dx.doi.org/10.1038/srep46450. DOI: 10.1038/srep46450.

[20] Coudray N, Ocampo P S, Sakellaropoulos T, et al. Classification and mutation prediction from non ‑ small cell lung cancer histopathology images using deep learning[J/OL]. Nature Medicine. 2018, 24(October):1. http://www.nature.com/articles/s41591-018-0177-5. DOI: 10.1038/s41591-018-0177-5.

[21] Habibzadeh Motlagh N, Jannesary M, Aboulkheyr H, et al. Breast Cancer Histopathological Image Classification: A Deep Learning Approach [M], 2018: 1–8. DOI: 10.1101/242818.

[22] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases[J]. Journal of Pathology Informatics. 2016, 7(1):29. DOI: 10.4103/2153-3539.186902.

[23] Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities[J]. Medical Image Analysis. 2016, 33:170–175. DOI: 10.1016/j.media.2016.06.037.

[24] Lin H, Chen H, Dou Q, et al. ScanNet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image[J]. Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018. 2018, 2018-Janua:539–546. DOI: 10.1109/WACV.2018.00065.

[25] Liu Y, Gadepalli K, Norouzi M, et al. Detecting Cancer Metastases on Gigapixel Pathology Images[M/OL]. 2017. http://arxiv.org/abs/1703.02442.

[26] Tellez D, Balkenhol M, Otte-Holler I, et al. Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks[J]. IEEE Transactions on Medical Imaging. 2018, 37(9):2126–2136. DOI: 10.1109/TMI.2018.2820199.

[27] Xu Y, Jia Z, Wang L B, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features[J/OL]. BMC Bioinformatics. 2017, 18(1):281. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1685-x. DOI: 10.1186/s12859-017-1685-x.

[28] Hosny A, van Griethuysen J J, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype[J]. Cancer Research. 2017, 77(21):e104–e107. DOI: 10.1158/0008-5472.can-17-0339.

[29] Chen P H, Gadepall K, MacDonald R, et al. An Augmented Reality Microscope for Real time Automated Detection of Cancer[M/OL]. 2018. https://research.googleblog.com/2018/04/an-augmented-reality-microscope.html.

[30] Cooper L A, Kong J, Gutman D A, et al. Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images[J/OL]. Laboratory Investigation. 2015, 95(4):366–376. http://dx.doi.org/10.1038/labinvest.2014.153. DOI: 10.1038/labinvest.2014.153.

[31] Yu K, Berry G J, Rubin D L, et al. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma Report Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma[J/OL]. Cell Systems. 2017, 1–8. https://doi.org/10.1016/j.cels.2017.10.014. DOI: 10.1016/j.cels.2017.10.014.

[32] LeCun Y, Bengio Y, Hinton G. Deep learning[J/OL]. Nature. 2015, 521(7553):436–444. http://www.nature.com/articles/nature14539. DOI: 10.1038/nature14539.

[33] Houghton J P, Ervine A J, Kenny S L, et al. Concordance between digital pathology and light microscopy in general surgical pathology: A pilot study of 100 cases[J]. Journal of Clinical Pathology. 2014, 67(12):1052–1055. DOI: 10.1136/jclinpath-2014-202491.

[34] Kurc T, Qi X, Wang D, et al. Scalable analysis of Big pathology image data cohorts using efficient methods and high-performance computing strategies[J/OL]. BMC Bioinformatics. 2015, 16(1):1–21. http://dx.doi.org/10.1186/s12859-015-0831-6. DOI: 10.1186/s12859-015-0831-6.

[35] Mercan C, Aksoy S, Mercan E, et al. Multi-Instance Multi-Label Learning for Multi-Class Classification of Whole Slide Breast Histopathology Images[J]. IEEE Transactions on Medical Imaging. 2018, 37(1):316–325. DOI: 10.1109/TMI.2017.2758580.

[36] Zhou N, Fedorov A, Fennessy F, et al. Large scale digital prostate pathology image analysis combining feature extraction and deep neural network [M], 2017: 1–14.

[37] El-Gabry E A, Parwani A V, Pantanowitz L. Whole-slide imaging: widening the scope of cytopathology[J/OL]. Diagnostic Histopathology. 2014, 20(12):456–461. http://linkinghub.elsevier.com/retrieve/pii/S1756231714001753. DOI: 10.1016/j.mpdhp.2014.10.006.

[38] Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project[J]. Nature Genetics. 2013, 45(6):580–585. DOI: 10.1038/ng.2653.

[39] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast CancerMachine Learning Detection of Breast Cancer Lymph Node MetastasesMachine Learning Detection of Breast Cancer Lymph Node Metastases[J/OL]. JAMA. 2017, 318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585. DOI: 10.1001/jama.2017.14585.

[40] Clunie D, Hosseinzadeh D, Wintell M, et al. Digital Imaging and Communications in Medicine Whole Slide Imaging Connectathon at Digital Pathology Association Pathology Visions 2017 [M], 2018: 1–4. DOI: 10.4103/jpi.jpi.

[41] Herrmann M D, Clunie D A, Fedorov A, et al. Implementing the DICOM Standard for Digital Pathology.[J/OL]. Journal of pathology informatics. 2018, 9(1):37. http://www.ncbi.nlm.nih.gov/pubmed/30533276http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6236926. DOI: 10.4103/jpi.jpi_42_18.

[42] Gutman D A, Cobb J, Somanna D, et al. Cancer digital slide archive: An informatics resource to support integrated in silico analysis of TCGA pathology data[J/OL]. Journal of the American Medical Informatics Association. 2013, 20(6):1091–1098. https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2012-001469. DOI: 10.1136/amiajnl-2012-001469.

[43] Schindelin J, Rueden C T, Hiner M C, et al. The ImageJ ecosystem: An open platform for biomedical image analysis[J]. Molecular Reproduction and Development. 2015, 82(7-8):518–529. DOI: 10.1002/mrd.22489.

[44] Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations[J]. International Journal of Computer Vision. 2017. DOI: 10.1007/s11263-016-0981-7.

[45] Dong F, Beck A, Irshad H, et al. Crowdsourcing Image Annotation for Nucleus Detection and Segmentation in Computational Pathology: Evaluating Experts, Automated Methods, and the Crowd [M], 2014: 294–305. DOI: 10.1142/9789814644730_0029.

[46] Xing F, Yang L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review[J]. IEEE Reviews in Biomedical Engineering. 2016, 9:234–263. DOI: 10.1109/RBME.2016.2515127.

[47] Achanta R, Shaji A, Smith K, et al. SLIC superpixels compared to state-of-the-art superpixel methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012, 34(11):2274–2281. DOI: 10.1109/TPAMI.2012.120.

[48] Kothari S, Phan J H, Stokes T H, et al. Pathology imaging informatics for quantitative analysis of whole-slide images[J]. Journal of the American Medical Informatics Association. 2013, 20(6):1099–1108. DOI: 10.1136/amiajnl-2012-001540.

[49] Delahunt B, Cheville J C, Martignoni G, et al. The International Society of Urological Pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters[J]. American Journal of Surgical Pathology. 2013, 37(10):1490–1504. DOI: 10.1097/PAS.0b013e318299f0fb.

[50] Sellaro T, Filkins R, Hoffman C, et al. Relationship between magnification and resolution in digital pathology systems[J/OL]. Journal of Pathology Informatics. 2013, 4(1):21. http://www.jpathinformatics.org/text.asp?2013/4/1/21/116866. DOI: 10.4103/2153-3539.116866.

[51] Zarella M D, Bowman D, Aeffner F, et al. A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association[J/OL]. Archives of Pathology & Laboratory Medicine. 2018, arpa.2018–0343–RA. http://www.archivesofpathology.org/doi/10.5858/arpa.2018-0343-RA. DOI: 10.5858/arpa.2018-0343-RA.

[52] Carney P A, Allison K H, Oster N V, et al. Identifying and processing the gap between perceived and actual agreement in breast pathology interpretation[J/OL]. Modern Pathology. 2016, 29(7):717–726. http://dx.doi.org/10.1038/modpathol.2016.62. DOI: 10.1038/modpathol.2016.62.

[53] Fuhrman S A, Lasky L C, Limas C. Prognostic significance of morphologic parameters in renal cell carcinoma[J]. American Journal of Surgical Pathology. 1982, 6(7):655–663.

[54] OLYMPUS CORPORATION. Resolving Power[M/OL]. https://www.olympus-ims.com/en/microscope/terms/resolving{_}power/.

[55] Rittscher J, Machiraju R, Wong S T C. Microscopic image analysis for life science applications [M]: Artech House, 2008: 489.

[56] Chen J, Xu J, Kang D, et al. Multiphoton microscopic imaging of histological sections without hematoxylin and eosin staining differentiates carcinoma in situ lesion from normal oesophagus[J]. Applied Physics Letters. 2013, 103(18):1–6. DOI: 10.1063/1.4826322.

[57] Elmore J G, Longton G M, Carney P A, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens[J]. JAMA - Journal of the American Medical Association. 2015, 313(11):1122–1132. DOI: 10.1001/jama.2015.1405.

[58] Delahunt B, Eble J N, Egevad L, et al. Grading of renal cell carcinoma[J]. Histopathology. 2019, 74(1):4–17. DOI: 10.1111/his.13735.

[59] Delahunt B, Egevad L, Samaratunga H, et al. Gleason and Fuhrman no longer make the grade[J]. Histopathology. 2016, 68(4):475–481. DOI: 10.1111/his.12803.

[60] Delahunt B, Sika-Paotonu D, Bethwaite P B, et al. Grading of clear cell renal cell carcinoma should be based on nucleolar prominence[J]. American Journal of Surgical Pathology. 2011, 35(8):1134–1139. DOI: 10.1097/PAS.0b013e318220697f.

[61] Delahunt B. Advances and controversies in grading and staging of renal cell carcinoma[J/OL]. Modern Pathology. 2009, 22(S2):S24–S36. http://dx.doi.org/10.1038/modpathol.2008.183. DOI: 10.1038/modpathol.2008.183.

[62] Marinelli R J, Montgomery K, Liu C L, et al. The stanford tissue microarray database[J]. Nucleic Acids Research. 2008, 36(SUPPL. 1):871–877. DOI: 10.1093/nar/gkm861.

[63] Verine J, Colin D, Nheb M, et al. Architectural Patterns are a Relevant Morphologic Grading System for Clear Cell Renal Cell Carcinoma Prognosis Assessment[J]. American Journal of Surgical Pathology. 2018, 42(4):423–441. DOI: 10.1097/PAS.0000000000001025.

[64] Yu K H, Zhang C, Berry G J, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features[J/OL]. Nature Communications. 2016, 7:12474. http://www.nature.com/doifinder/10.1038/ncomms12474. DOI: 10.1038/ncomms12474.

[65] Nir G, Karimi D, Goldenberg S L, et al. Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images[J]. JAMA network open. 2019, 2(3):e190442. DOI: 10.1001/jamanetworkopen.2019.0442.

[66] Yeh F C, Parwani A V, Pantanowitz L, et al. Automated grading of renal cell carcinoma using whole slide imaging.[J/OL]. Journal of pathology informatics. 2014, 5(1): 23. http://www.ncbi.nlm.nih.gov/pubmed/25191622http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4141422. DOI: 10.4103/2153-3539.137726.

[67] Kim G, Rajasekaran S A, Thomas G, et al. Renal clear-cell carcinoma: An ultrastructural study on the junctional complexes[J]. Histology and Histopathology. 2005, 20(1):35–44. DOI: 10.14670/HH-20.35.

[68] Yu K, Berry G J, Rubin D L, et al. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma[J/OL]. Cell Systems. 2017, 5(6):620–627.e3. https://www.sciencedirect.com/science/article/pii/S2405471217304842. DOI: 10.1016/J.CELS.2017.10.014.

[69] Mukhopadhyay S, Feldman M D, Abels E, et al. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology of 1992 Cases (Pivotal Study)[J]. American Journal of Surgical Pathology. 2018, 42(1):39–52. DOI: 10.1097/PAS.0000000000000948.

[70] Jones A D, Graff J P, Darrow M, et al. Impact of pre-analytic variables on deep learning accuracy in histopathology[J]. Histopathology. 2019, 0–3. DOI: 10.1111/his.13844.

[71] Helin H, Tolonen T, Ylinen O, et al. Optimized JPEG 2000 compression for efficient storage of histopathological whole-Slide images[J/OL]. Journal of Pathology Informatics. 2018, 9(1): 20. http://www.jpathinformatics.org/article.asp?issn=2153-3539year=2018volume=9issue=1spage=20epage=20aulast=Helin. DOI: 10.4103/jpi.jpi_69_17.

[72] Kalinski T, Zwönitzer R, Grabellus F, et al. Lossless compression of JPEG2000 whole slide images is not required for diagnostic virtual microscopy[J]. American Journal of Clinical Pathology. 2011, 136 (6):889–895. DOI: 10.1309/AJCPYI1Z3TGGAIEP.

[73] Microsoft. Deep Zoom[M/OL]. 2011. https://docs.microsoft.com/en-us/previous-versions/windows/silverlight/dotnet-windows-silverlight/cc645050{%}28v{%}3Dvs.95{%}29.

[74] Singh R, Chubb L, Pantanowitz L, et al. Standardization in digital pathology: Supplement 145 of the DICOM standards[J/OL]. Journal of pathology informatics. 2011, 2:23. https://www.ncbi.nlm.nih.gov/pubmed/21633489https://www.ncbi.nlm.nih.gov/pmc/PMC3097525/. DOI: 10.4103/2153-3539.80719.

Appendix A MySQL Data Model

A.1 Codes

Code A-1 Code for OpenHI data model construction

```
-- Create the schema version 4

CREATE DATABASE `db_wsi4-1_dev` ; -- OR `db_wsi1`

USE `db_wsi4-1_dev`;

CREATE TABLE `annotator` (
  `annotator_id` smallint (5) unsigned NOT NULL AUTO_INCREMENT,
  `password` varchar (45) COLLATE utf8_unicode_ci DEFAULT 'NA',
  PRIMARY KEY (`annotator_id`)
) ENGINE=InnoDB ;

-- Insert annotator default password (2 annotators )
INSERT INTO annotator(password) values ('123456');
INSERT INTO annotator(password) values ('123456');

CREATE TABLE `patient` (
  `patient_id` int (10) unsigned NOT NULL AUTO_INCREMENT,
  `tcga_case_id` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `patient_info` JSON,
  PRIMARY KEY (`patient_id`),
  INDEX(tcga_case_id)
) ENGINE=InnoDB ;

CREATE TABLE `biospecimen` (
  `bio_id` int (10) unsigned NOT NULL AUTO_INCREMENT,
  `tcga_case_id` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `bio_info` JSON,-- end editing
  PRIMARY KEY (`bio_id`),
  CONSTRAINT `tcga_case_id_in_bio` FOREIGN KEY (`tcga_case_id`) REFERENCES
      `patient` (`tcga_case_id`) ON DELETE CASCADE ON UPDATE CASCADE
) ENGINE=InnoDB ;

CREATE TABLE `wsi` (
  `slide_id` smallint (6) unsigned NOT NULL AUTO_INCREMENT,
  `tcga_wsi_id` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `tcga_case_id` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `tcga_wsi_slide_id` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `uuid` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
  `filename` varchar (255) COLLATE utf8_unicode_ci DEFAULT 'NA',
#   `bio_id` int (10) unsigned NOT NULL,
  PRIMARY KEY (`slide_id`),
```

```
42          CONSTRAINT `tcga_case_id_in_wsi` FOREIGN KEY (`tcga_case_id`) REFERENCES
                `patient` (`tcga_case_id`) ON DELETE CASCADE ON UPDATE CASCADE
43      #   CONSTRAINT `bio_id_in_wsi` FOREIGN KEY (`bio_id`) REFERENCES `biospecimen` (
            `bio_id`) ON DELETE CASCADE ON UPDATE CASCADE
44      ) ENGINE=InnoDB AUTO_INCREMENT=1 ;
45
46      # -- Insert mock-up wsi list with a loop
47      # drop procedure if exists load_foo_test_data ;
48      #
49      # USE `db_wsi4-1_dev`;
50      # delimiter #
51      # create procedure load_foo_test_data ()
52      # begin
53      #
54      # declare v_max int unsigned default 1000;
55      # declare v_counter int unsigned default 0;
56      #   while v_counter < v_max do
57      #     INSERT INTO `db_wsi4-1_dev`.`wsi` (`tcga_wsi_id`, `tcga_case_id` , `uuid` )
            VALUES ('123456', '123', '123');
58      #     set v_counter=v_counter+1;
59      #   end while;
60      #   commit;
61      # end #
62      #
63      # delimiter ;
64      #
65      # call load_foo_test_data () ;
66
67
68      CREATE TABLE `grading` (
69        `grading_id` tinyint (3) unsigned NOT NULL AUTO_INCREMENT,
70        `grading_std_name` varchar(255) COLLATE utf8_unicode_ci DEFAULT 'NA',
71        PRIMARY KEY (`grading_id`)
72      ) ENGINE=InnoDB ;
73
74      -- Insert ISUP grading system
75      INSERT INTO `db_wsi4-1_dev`.`grading` (`grading_std_name`) VALUES ('ISUP1');
76      INSERT INTO `db_wsi4-1_dev`.`grading` (`grading_std_name`) VALUES ('ISUP2');
77      INSERT INTO `db_wsi4-1_dev`.`grading` (`grading_std_name`) VALUES ('ISUP3');
78      INSERT INTO `db_wsi4-1_dev`.`grading` (`grading_std_name`) VALUES ('ISUP4');
79
80      CREATE TABLE `pslv` (
81        `pslv_id` tinyint (3) unsigned NOT NULL AUTO_INCREMENT,
82        `subregion_density` int (10) unsigned DEFAULT NULL,
83        PRIMARY KEY (`pslv_id`)
84      ) ENGINE=InnoDB ;
85
86      -- Insert sub-region density
87      INSERT INTO `db_wsi4-1_dev`.`pslv` (`subregion_density`) VALUES (5000);
88      INSERT INTO `db_wsi4-1_dev`.`pslv` (`subregion_density`) VALUES (1000);
```

```
89  INSERT INTO `db_wsi4-1_dev`.`pslv` (`subregion_density`) VALUES (60);
90
91  CREATE TABLE `point` (
92    `pt_id` int(10) unsigned NOT NULL AUTO_INCREMENT,
93    `x` int(10) unsigned NOT NULL,
94    `y` int(10) unsigned DEFAULT NULL,
95    `annotation_ts` datetime NOT NULL, -- Annotation timestamp
96    `grading_id` tinyint(3) unsigned NOT NULL,
97    `slide_id` smallint(6) unsigned NOT NULL,
98    `annotator_id` smallint(5) unsigned NOT NULL,
99    `pslv_id` tinyint(3) unsigned DEFAULT NULL,
100   -- keep adding code
101   `region_id` int(10) unsigned NOT NULL,
102   `selected` tinyint(2) unsigned NOT NULL,
103   `anno_batch` int(10) unsigned NOT NULL,
104   -- end editing
105   PRIMARY KEY (`pt_id`),
106   KEY `INDEX` (`x`,`y`) USING BTREE,
107   KEY `annotator_id_idx` (`annotator_id`),
108   KEY `slide_id_idx` (`slide_id`),
109   KEY `grading_id_idx` (`grading_id`),
110   KEY `pslv_id_idx` (`pslv_id`),
111   CONSTRAINT `annotator_id` FOREIGN KEY (`annotator_id`) REFERENCES `annotator` (
        `annotator_id`) ON DELETE CASCADE ON UPDATE CASCADE,
112   CONSTRAINT `grading_id` FOREIGN KEY (`grading_id`) REFERENCES `grading` (
        `grading_id`) ON DELETE CASCADE ON UPDATE CASCADE,
113   CONSTRAINT `pslv_id` FOREIGN KEY (`pslv_id`) REFERENCES `pslv` (`pslv_id`) ON
        DELETE CASCADE ON UPDATE CASCADE,
114   CONSTRAINT `slide_id` FOREIGN KEY (`slide_id`) REFERENCES `wsi` (`slide_id`) ON
        DELETE CASCADE ON UPDATE CASCADE
115 ) ENGINE=InnoDB ;
116
117 -- To be annotated (tba) list support
118 CREATE TABLE `tba_list` (
119   `sw_id` int(10) unsigned NOT NULL AUTO_INCREMENT,
120   `slide_id` smallint(6) unsigned NOT NULL,
121   `center_x` int(10) unsigned NOT NULL,
122   `center_y` int(10) unsigned NOT NULL,
123   PRIMARY KEY (`sw_id`),
124   KEY `slide_id_idx` (`slide_id`),
125   CONSTRAINT `slide_id_in_tba` FOREIGN KEY (`slide_id`) REFERENCES `wsi` (`slide_id`)
        ON DELETE CASCADE ON UPDATE CASCADE
126 ) ENGINE=InnoDB ;
127
128 CREATE TABLE `annotator_sw_mark` (
129   `mark_id` int(10) unsigned NOT NULL AUTO_INCREMENT,
130   `sw_id` int(10) unsigned NOT NULL,
131   `annotator_id` smallint(5) unsigned NOT NULL,
132   PRIMARY KEY (`mark_id`),
133   KEY `sw_id_idx` (`sw_id`),
```

```
134             KEY ‘annotator_id_idx‘  ( ‘annotator_id‘ ) ,
135             CONSTRAINT ‘sw_id_in_mark‘ FOREIGN KEY (‘sw_id‘) REFERENCES ‘tba_list‘ (‘sw_id‘)
                    ON DELETE CASCADE ON UPDATE CASCADE,
136             CONSTRAINT ‘annotator_id_in_mark‘ FOREIGN KEY (‘annotator_id‘) REFERENCES
                    ‘annotator‘ (‘annotator_id‘) ON DELETE CASCADE ON UPDATE CASCADE
137           ) ENGINE=InnoDB ;
```

Appendix B Image of ccRCC Nuclei Samples

WHO/ISUP Grade 1:

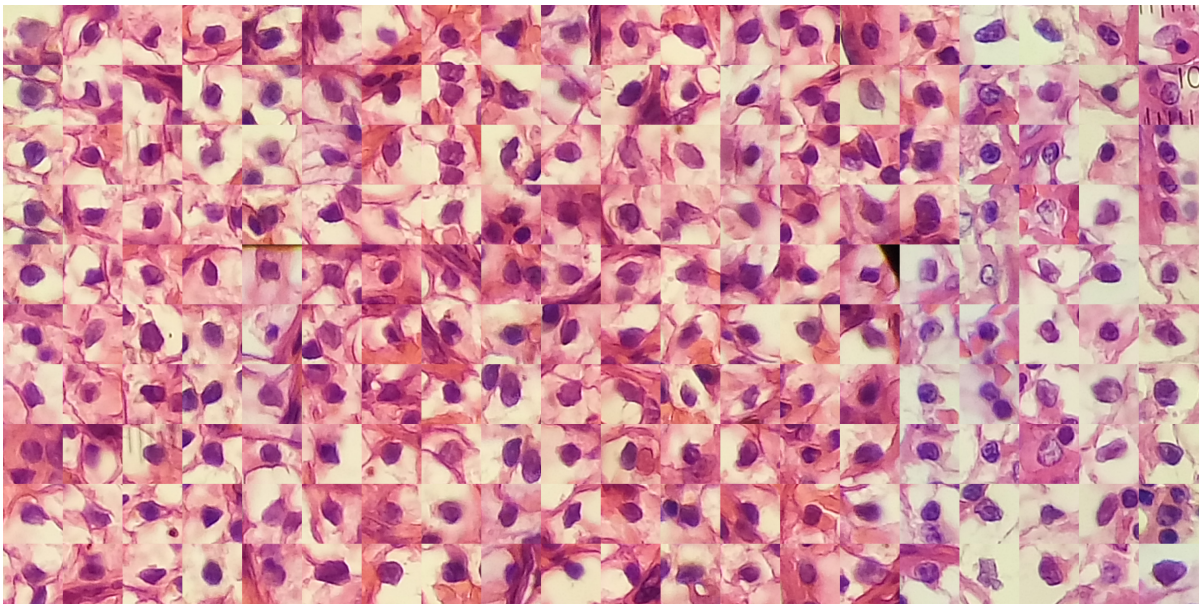


Figure B-1 Graded 1 nuclei according to WHO/ISUP grading standard fro clear cell renal cell carcinoma.

WHO/ISUP Grade 2:

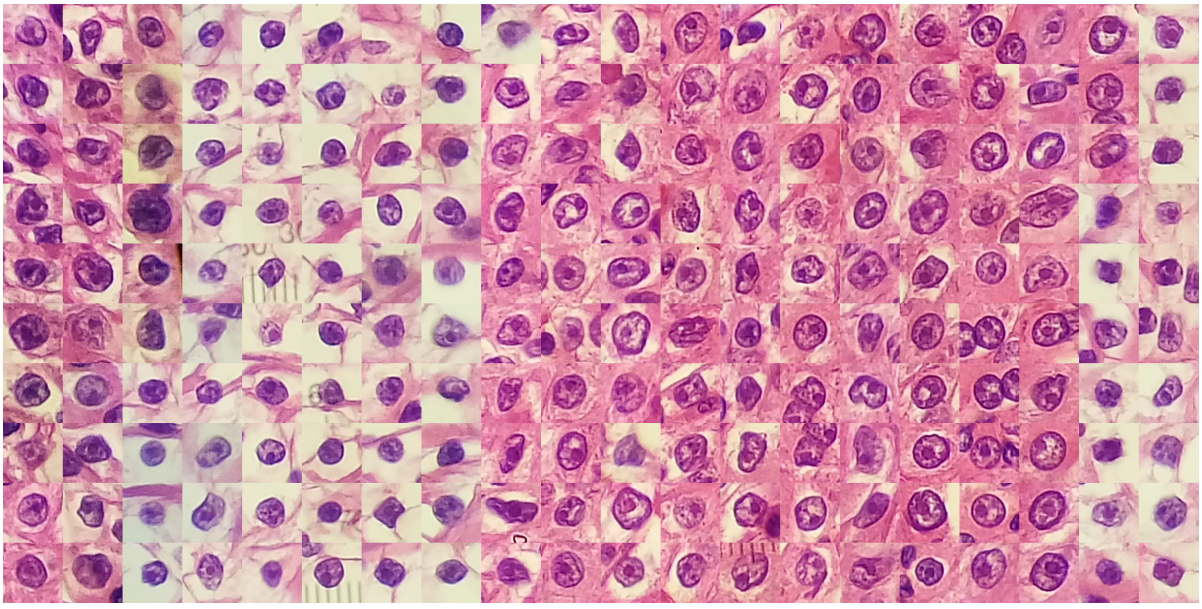


Figure B-2 Graded 2 nuclei according to WHO/ISUP grading standard fro clear cell renal cell carcinoma.

WHO/ISUP Grade 3:

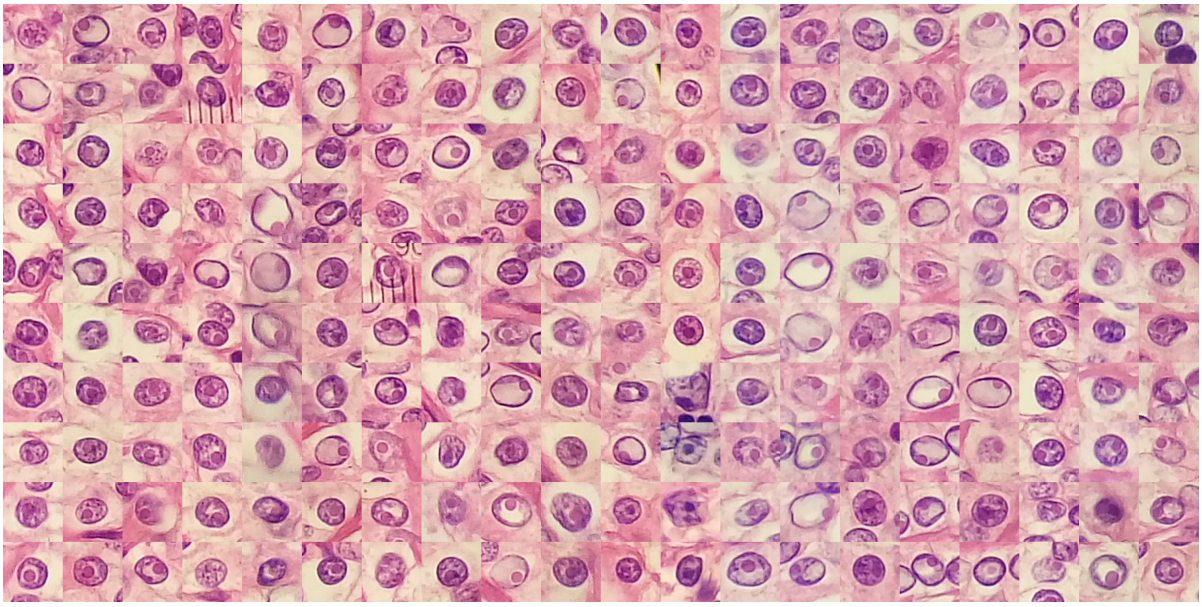


Figure B-3 Graded 3 nuclei according to WHO/ISUP grading standard fro clear cell renal cell carcinoma.

Achievements

Publications：

[1] Puttapirat P, Zhang H, Deng J, et al. OpenHI: Open platform for histopathological image annotation[J]. International Journal of Data Mining and Bioinformatics, 2019, X(Y): xxxx. (accepted on 2019.05.22)

[2] Dong Y, Puttapirat P, Deng J, et al. LibMI - an open source library to efficiently read, modify and write extremely large images[J]. Scientific Data, 2019, XX(XX): XX. (under review on 2019.05.20)

[3] Puttapirat P, Zhang H, Lian Y, et al. OpenHI - An open source framework for annotating histopathological image[C]. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE, 2018: 1076–1082.

Patent:

[1] Li C, Puttapirat P, Zhang H. 基于云端的大型病理学图像协作注释方法及系统: China, 201910252955.9[P]. China: 2019.

学位论文独创性声明 (1)

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 交回学校授予的学位证书；
2. 学校可在相关媒体上对作者本人的行为进行通报；
3. 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉；
4. 本人负责因论文成果不实产生的法律纠纷。

论文作者 (签名)：　　日期：　　年　　月　　日

学位论文独创性声明 (2)

本人声明：研究生_____ 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 学校可在相关媒体上对本人的失察行为进行通报；
2. 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉；
3. 本人接受学校按照有关规定做出的任何处理。

指导教师 (签名)：　　日期：　　年　　月　　日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者 (签名)：　　日期：　　年　　月　　日

指导教师 (签名)：　　日期：　　年　　月　　日