Recognizing Magnification Levels in Microscopic Snapshots using Machine Learning

by

Manit Maulinkumar Zaveri

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.
I understand that my thesis may be made electronically available to the public.

Statement of Contribution

This thesis is partly based on the following papers that were authored by or co-authored by me:

A. **Zaveri, M.**, Kalra, S., Babaie, M., Shah, S., Damskinos, S., Kashani, H. and Tizhoosh, H.R., 2020. Recognizing Magnification Levels in Microscopic Snapshots. arXiv preprint arXiv:2005.03748.

B. **Zaveri, M.**, Hemati, S., Damskinos, S., Shah, S., Tizhoosh, H. Kimia-5MAG – A Dataset for Learning the Magnification in Histopathology Images. Submitted to IEEE's 32^{nd} International Conference on Tools with Artificial Intelligence, July 2020.

C. Riasatian, A., Babaie, M., Maleki, D., Kalra, S., Valipour, M., Hemati, S., **Zaveri, M**, et al. KimiaNet – Fine-Tuning and Training of DenseNet for Histopathology Image Representation Using TCGA Diagnostic Slides. Submitted to Elsevier's Medical Image Analysis, June 2020.

I am the first author, of Papers A and B, and I contributed to their implementation, experimentation, and writing. My contributions for Paper C were to a lesser extent. A major part of the experimental discussion and results in chapter 3 and chapter 4 are taken from Papers A and B. Paper A has been accepted for publication in IEEE EMBC'2020, Papers B has been accepted for publication in IEEE ICTAI'2020, and Paper C is under review.

Abstract

State-of-of-the-art computer vision research has facilitated technology evolution in the field of medical imaging. The primary achievement of the imaging algorithms developed is the extraction of expressive features from digital images. The real advantage of this progress can be observed when these features are utilized in the primary tasks of Content Based Image Retrieval (CBIR) and developing tissue classification systems for confirming the diagnostic results of medical images. Digital Pathology (DP) is a branch of medical imaging focused on digital images acquired from histopathology specimens (i.e., biopsy samples). The camera-mounted microscope was introduced in the late 1960s, and is one of the most popular, convenient, and effective tool to generate a digital footprint of tissues from glass slides. The introduction of Whole Slide Imaging (WSI) technology has completely changed Light Microscopy (LM). Existing datasets that are modeled using microscopic camera systems are given the least attention by the research community because of i) missing relative information, i.e., magnification level and annotations, ii) low-resolution images, and iii) easily accessible WSI slides or patches with relevant information.

With the increasing demand for accurate diagnosis of diseases such as cancer, there is an imminent need to utilize the knowledge not only from WSI images but also from microscopic snapshots by using state-of-the-art Machine Learning (ML) techniques to meet the pressing demand for more reliable diagnosis. This thesis is an empirical study to investigate methods for recognizing the magnification level of microscopic images to enable their application in various tasks. Additional investigations to understand the influence of the primary site (i.e., organs) on recognizing magnification levels were conducted. Qualitative assessments of feature extraction algorithms, such as Local Binary Pattern (LBP), and several pre-trained Convolution Neural Network (CNN) architectures are provided. These algorithms are used as feature extractors to comprehend knowledge at an individual magnification level from microscopic snapshots of histopathology images. The classification is conducted by three traditional classifiers, Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (K-NN), by implementing traditional computer vision and Deep Learning (DL) algorithms to learn the magnification level associated with each microscopic snapshot. Three different datasets were used to conduct the experiments, which were evaluated using total accuracy, patient or primary site accuracy, and F1-score. The total accuracies and F1 scores were 93.26% and 0.94 for the *KIMIA-MAG-5* dataset, 91.50% and 0.93 for the *BreakHis* dataset, and 87.11% and 0.87 for the *OMAX* dataset, respectively. An insight from the primary site analysis shows that the task of recognizing magnification levels in images of pleura, lungs, and breasts are straightforward.

Acknowledgements

Writing this thesis has been a fascinating and extremely rewarding experience. I would like to thank all those who have contributed to my thesis in different ways.

First and foremost, I would like to thank my supervisor, Professor Hamid R. Tizhoosh, for allowing me to continue my academic endeavors by pursuing an M.Sc. degree under his supervision. His constant support, encouragement, and patience will always be appreciated. A quick chat with him gives one the power they need to accomplish remarkable milestones. I am glad to be a part of his lab, Kimia Lab, where all researchers are free to explore creative ideas.

Special thanks to Professor Jonathan Kofman and Professor Siby Samuel for taking the time to review my thesis and provide valuable suggestions.

I would also like to acknowledge the scholarship program offered by the Government of Canada (NSERC) for providing much-needed financial support throughout my degree.

I would like to thank all my co-authors, Shivam Kalra, Sobhan Hemati, Morteza Babaie, Savvas Damaskinos, and Sultaan Shah, lab members, Amir Safarpoor and Milad Sikroudi, and others for accompanying me throughout this journey.

I appreciate the collaboration of Huron Digital Pathology for providing tissue specimens to curate the OMAX dataset. I owe my special thanks to Dr. Morteza Babaie and Mahajhabin Sajadi for capturing the images.

I would like to thank Shivam Kalra for his help in managing the in-house server and virtual database. Inaccessibility to the Isilon server and cloud computing would restrict me from experimenting on such diverse datasets.

I would like to thank my parents, Maulin and Hetal Zaveri for raising me to value education. I deeply appreciate my sister Saloni Zaveri for all her support, as my endeavors at the University of Waterloo are solely because of her. Special thanks to Shim Patel, Yash Shah, Jeet Shah, and all my friends for being supportive and motivating me throughout my academic journey.

Finally, I pay my respects to God, for bestowing me with this opportunity to pursue my education from the most renowned university, and for all the good health, courage, inspiration, and light.

Dedication

This is dedicated to my beloved parents and sister.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence x, 1–3, 6, 7, 9, 16, 26

CAD Computer Assisted Diagnosis 2, 10, 15, 16, 54

CBIR Content Based Image Retrieval iv, x, 3, 6, 7, 17

CNN Convolution Neural Network iv, 7, 8, 16, 17, 21–23, 39–41, 43, 53

DL Deep Learning iv, 6, 8, 16, 17, 21, 22, 26, 54

DP Digital Pathology iv, 2, 3, 5, 7, 15, 16, 55

K-NN K-Nearest Neighbors iv, 8, 40, 47, 48, 53, 54

LBP Local Binary Pattern iv, xiii, 8, 18, 20, 40, 41, 53

LM Light Microscopy iv, 3, 14

ML Machine Learning iv, 1–3, 15, 16, 26

RF Random Forest iv, 8, 25, 40, 41, 47, 48, 53, 54

SVM Support Vector Machine iv, 8, 25, 40–42, 44, 46–48, 53, 54

WSI Whole Slide Imaging iv, 2, 6, 13, 17

Chapter 1

Introduction

"As to diseases, make a habit of two things to help or at least to do no harm."

– Hippocrates

"DIAGNOSING cancer at a later stage and the inability to provide treatment condemns many people to unnecessary suffering and early death" [1]. A report published by World Health Organization in early 2017 titled "Early cancer diagnosis saves lives, cuts treatment costs" focuses mainly on developing guidelines for diagnosing cancer at an early stage, which significantly increases the chances of survival by reducing disease severity and eventually saving treatment costs. By 2030, we will have an estimated 21 million cases of cancer worldwide. Some of these cases can be prevented from becoming fatal by implementing proper preventive measures and prescreening routines, resulting in early diagnosis. The main purpose of the guidelines published by health agencies is to serve as a baseline for providing information to certain parts of society that are at risk of being infected in the future. *Overdiagnosis* is the term used for asymptomatic, harmless diagnosis during preventive screening. This is a primary source of controversy in science and has created a movement against prescreening in the last decade. This debate provides a platform for the research and application of this study by helping automated cancer diagnosis systems to assess the tradeoff between early diagnosis and overdiagnosis using state-of-the-art computational algorithms [2].

The success of Artificial Intelligence (AI) and Machine Learning (ML) in practice has encouraged researchers to implement automated algorithms for clinical applications by

integrating computers with traditional healthcare practices. The increasing applications of these algorithms in medicine with field experts has proved to be beneficial. Healthcare systems are generally considered rather slow in keeping pace with technology because of their sensitive nature in dealing with human life.[1]

The primary motivation for this study is to contribute to alleviating the digitization of DP and its applications such that future demands are met efficiently. In this study, we mainly focus on developing a framework composed of AI algorithms to extract the features of pathology images collected using two different modalities and predicting an important characteristic prevalent in the histopathology domain—magnification level. The main purpose of this system is to learn the unknown magnification level of images generated using a camera-based microscope such that many existing repositories become usable for learning. This will allow more data to flow into the pipeline of research and development of computer-assisted systems in clinical practice.

1.1 Problem Statement

In this study, we mainly aim to create a universal pre-processing method to contribute toward more realistic and optimal Computer Assisted Diagnosis (CAD) systems capable of safely assisting multi-magnification diagnosis of histopathology images. The proposed feature will directly impact the results of all classification and image search systems being deployed in clinical practice; thus, allowing improved accuracy for multi-magnification query images.

The main problem that hinders the deployment of ML algorithms in healthcare is the lack of well-annotated and -labeled data. In the last decade, the popularity of publicly available datasets composed of WSI virtual slides or patches has increased [3, 4]. In the same time, pathology has seen a paradigm shift from traditional microscopic camera images to high-resolution images with exceptional visual clarity and digital magnification attributes for viewing specimens on a computer screen. As a result of these advantages, WSI technology has a major limitation of high memory consumption; thus, limiting its usage in the development of AI-powered systems because it makes their applications complex at a high computational cost. Currently, the majority of pathologists still use camera-based microscopes because of their simplicity and wide availability. As a result, an enormous number of microscopic snapshots have been acquired for various reasons but not considered for automated processing because of their low resolution, uncatalogued magnification

[1] https://tinyurl.com/y8kgrk9c

levels, and missing related information. Another significant factor that hinders the universal application of ML algorithms is the variations in image resolution due to divergent image-capturing modalities(i.e., camera mounted microscope and whole slide scanner). As a result, the shortage of balanced datasets with diversity in magnification level, site of origin, and type of staining adds to the motivation for this work.

Human error is considered to be the primary source of error in clinical practice. The majority of human errors are caused by limitations in human perception, leading to errors such as variations in data preparation, manual errors during capturing or scanning slides, and tissue folding while preparing glass slides [5]. The tediousness of pathology routines may negatively affect the decision-making capability of pathologists. Under severe daily stress, pathologists are also responsible for writing detailed reports explaining the nature of tissue, area of origin, visual characteristics, and diagnosis. Pathology reports are not standardized, meaning important information of the magnification at which the specimen slide was observed is lost. This magnification level is an important factor for two reasons: i) for the research community to develop magnification-invariant processing systems and ii) to obtain similar images at the same magnification level using an advanced CBIR system. In Light Microscopy (LM), the enlarged view of the glass slides allows the pathologist to zoom in/out, enabling better inspection of infinitesimal structures of cells, morphology, and structural patterns necessary for diagnosis. This specific area is in itself a challenge for DP. Currently, all digital solutions in practice are trained at specific magnification; thus, existing solutions focus on a specific task, such as tissue classification or image search at static magnifications. This limitation may have consequences for the wide adoption of such methods.

As the number of cancer cases increases, it becomes obligatory to invest more knowledge in exploring the potential of AI in histopathology. The principal challenge in this study is to build a universal AI-enabled pipeline that can classify the magnification level by understanding the vivid structural patterns observed at diverse magnification levels inspected by pathologists. This system can be considered as a by-default pre-processing step for all AI solutions deployed in the future when microscopic snapshots are used, ensuring more confident interpretation in CBIR and tissue classification systems of the future.

1.2 Background

Human research is extensive in terms of accessibility. Available resources range from exploring planetary movements at several yottameters to sub-atomic level changes in the

nucleus and cells in the range of zeptometers [2].

"Histology" and "pathology" are individual branches of biomedical science. Histology is the science of microscopic evaluation, and pathology is the study of the causes and effects of disease through laboratory evaluation of body tissues [3]. When the principles of pathology are combined with those of histology, the domain of "histopathology" emerges as the study of diagnosis of diseases by evaluating tissue samples obtained through biopsy using microscopes to examine the underlying tissue patterns [6].

Pathologists are domain experts in implementing pathology practices. Histopathology is a pivotal part of all healthcare systems. The primary task of a pathologist is microscopic analysis of tissues prepared on glass slides [7]. A detailed report composed of all necessary findings from a clinical evaluation is prepared by a pathologist [8]. These reports are the basis for many clinical decisions and are usually accompanied by a patient's medical history, images related to the current diagnosis, documentation of diagnosis, physical findings from the examinations, and a plan of treatment and treatment responses. These reports are a crucial part of medical history, and hence must be comprehensive and reliable.

Figure 1.1 shows the general complexity of image magnification. Image zooming is an important factor in the pathology domain as well as in satellite imaging and astrophysics. Image contents become more distinct with increasing magnification power, resulting in final images for both scenarios (i. e., satellite and pathology), that provide a clear insight into the minute components of images at maximum magnification power.

1.2.1 Influence of Digitization on Light Microscopy

In conventional microscopy, a pathologist changes the objective lens of the microscope to obtain an enhanced (field of) view, facilitating better access to traits of the image that cannot be seen with the naked eye. Microscopes are the only optical instrument that allow live image magnification for histology images, allowing interpretation of cells and tissues at high resolution [9]. The routine procedure conducted by pathologists is to repeatedly snap through multiple magnifications to grasp the differentiating characteristics of the specimen between varying scales of the objective lens. The total magnification of the microscope system is composed of the powers of the eyepiece and objective lens. While the magnification power of the eyepiece is fixed at 10x, the pathologist has a range of powers from 2.5x to 2000x of objective lenses [10].

[2] https://en.wikipedia.org/wiki/Orders_of_magnitude_(length)
[3] https://www.rcpath.org/discover-pathology/news/fact-sheets/histopathology.html

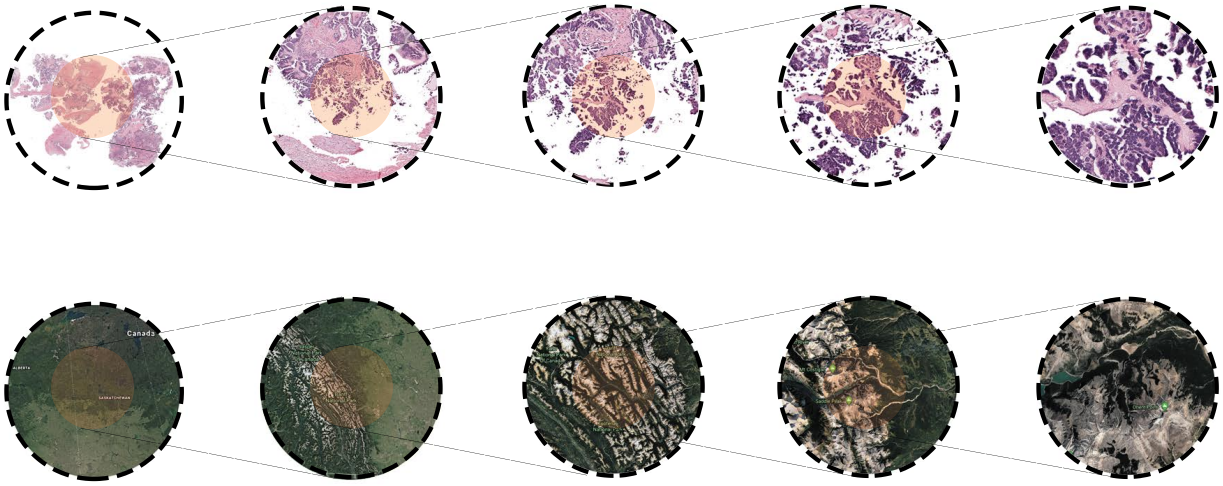


Figure 1.1: Illustrative example of microscope functionality of inter-changeable lens compared to satellite images. From left to right: (top) The smallest magnification level image is not able to distinguish fine-grained image components, capturing an overview of the tissue and cellular matrix. With increasing magnification power, the clear structure of the nuclei as well as the boundary of the cellular matrix become visible. (bottom) The leftmost image from Google Maps focuses on terrain in Canada. By increasing magnification power, distinct patterns of mountains, valleys, and waterbodies are observed, resembling the satellite view of Banff National Park in the rightmost image.

With the influence of developments achieved in the "digital age", DP can be defined as "integration of digital systems into conventional light microscopy, including sharing, monitoring, and interpreting pathology materials" [11]. Because of the traditional practice and popularity of microscopes among pathologists, they tend to adhere to the conventional techniques for clinical practice, even after the integration of a digital environment. Before the last decade, the technique that created the most significant impact in pathology was camera-mounted microscope systems, allowing pathologists to conduct their clinical work on microscopes while saving images in digital form. A simple camera attachment was added on top of the eyepiece to save a digital copy of an image at the same magnification as viewed by the expert. However, this coordination has its limitations: i) Various laboratories use different cameras as the add-on camera to capture snapshots, producing images with various colors, lighting, and resolutions. ii) Image quality may not be acceptable to the research community because of low resolution. iii) Associated information, such as the

magnification level of the captured image, is generally not available for all images. These problems limit the efficient use of enormous repositories of snapshots, making them the least used in medical research. With the introduction of whole-slide digital scanners in the last decade, all three problems were solved, and the research community has started developing and training AI solutions using this high-end data. The major limitation of whole-slide scanners and their implementation is their functional requirement of complex hardware and economic liability for installation and set-up. The cost of a whole-slide scanner compared to a camera mounted on a microscope is very high [12]. WSI technology represents the actual laboratory setting of a microscope in a virtual form on digital screens using software systems [13].

AI has sparked a revolution in digital pathology. With increasing data collection and proven stability of some AI algorithms, solutions powered with such algorithms have been witnessed in the real world [14, 15]. For example, a complete system for tissue-based diagnosis was developed using the power of AI in digital pathology [16]. With increasing applications of DL, we can achieve human-like performance using computerized diagnosis. Automated cancer classification is currently one of the most popular areas of research due to the success achieved by AI systems. The best example of proven success is in diagnosing tissue malignancy from breast muscles [17]. These prominent results have been accomplished solely because of increased data availability for this particular type of cancer. Particularly for breast cancer, DL computerized systems may even outperform human pathologists in identifying tumor metastasis [18]. Another application of DL, which has a significant impact on the healthcare community, is Content Based Image Retrieval systems, which can find similar types of images to support the decisions of pathologists, allowing a secondary assessment for the same slide [19]. This is achieved by several methods to match the similarity of the characteristics of a new query image with images from a database composed of several thousand plus images containing similar image characteristics. Figure 1.2 illustrates the traditional pathology environment with important steps that greatly affect the quality of diagnosis. The extension of AI systems to assist at every step is feasible.

1.3 Contribution of Thesis

The primary contribution of this thesis is an empirical study of frameworks for classifying the magnification levels of pathology images. Experiments were conducted to provide a comprehensive analysis of feature extraction algorithms and transfer learning using DenseNet121, ResNet151, Inception, EfficientNetB0, and VGG for their abilities to quantify image features from microscopic snapshots and their suitability for classification of the magnification level based on the tissue orientation. This study contributes to the goal of

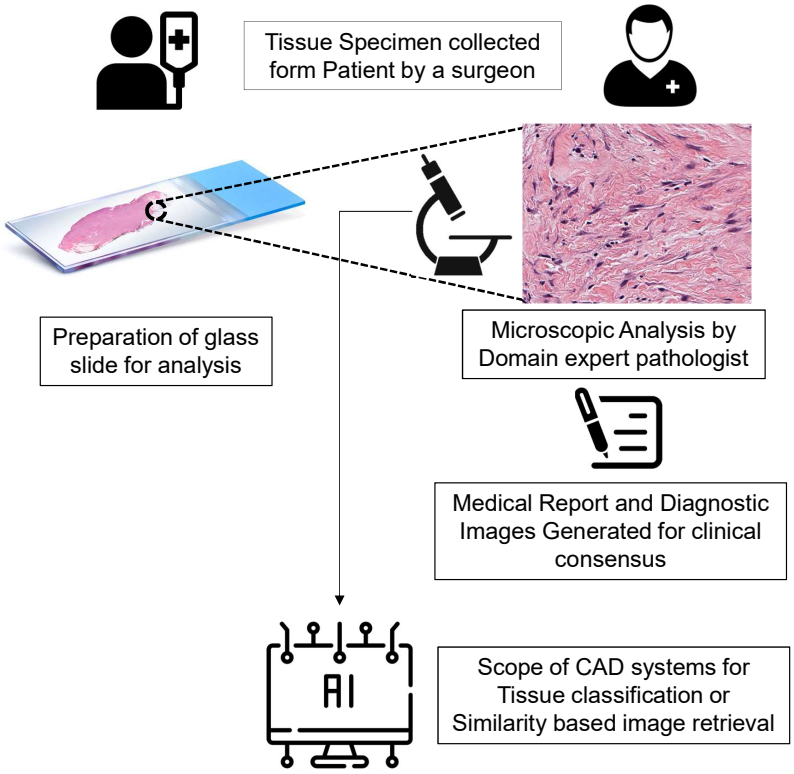


Figure 1.2: Complete overview of traditional microscopy, including steps of biopsy collection from the organ under investigation, glass examination under microscope using the prepared tissue specimen, making a clinical evaluation, and scope of AI systems for tasks of CBIR and automated cancer detection.

the biomedical community to integrate AI systems with healthcare devices. For universal compliance of AI in pathology, computer algorithms must learn the interpretation of image morphology similarly to pathologists, at multiple magnifications. This study directly addresses the problem of limited data in the research workflow. By classifying magnification levels of histopathology images, a well-balanced dataset can be created, which can eventually be used to develop capable solutions. Currently, most CNN models implemented in DP are trained on natural scenic images. Based on the experimental results, a pre-processing system for classifying pathology images as per their magnification level before training a solution for specific diagnosis or obtaining similar images using CBIR systems is proposed. In this thesis, feature extraction methods using traditional computer

vision algorithms with deep embeddings available using high-end DL techniques are compared. In this study, we conduct quantitative analysis on three datasets using feature extraction algorithms, further classifying them for recognizing magnification levels. In depth investigations are conducted for individual magnification levels and primary sites. The performances of handcrafted feature extraction methods and pre-trained CNNs as feature extractors are evaluated by classifying image features from the i) KIMIA-MAG-5, ii) Breakhis, and iii) OMAX datasets using supervised machine learning algorithms. RF, SVM, and K-NN classification algorithms are tested for their classification power to identify diverse patterns extracted using various feature extraction algorithms. Experimental results show that all the techniques are suitable for extracting important features from images. Parameter tuning for each classifier proves to beneficial for extracted features, making them more sensitive to minute differences. Additionally, the results of handcrafted features on KIMIA-MAG-5 show that Local Binary Pattern, a traditional computer vision algorithm, performs fairly well but is incomparable to the use of deep embeddings. DL offers more flexibility to adjust the network architectures; the exemplary success of pre-trained CNNs can be seen by conducting rigorous experiments. Therefore, in this study, we aim to answer the following questions:

- Is magnification recognition of microscopic images possible?
- Which methodology is suitable for this task?
- Which magnification levels are easy/difficult to predict?
- Can we recognize magnification levels across all internal organs?
- What are the prime factors in misclassified images?

An empirical study is conducted by implementing multiple techniques on three datasets to answer these questions.

Chapter 2

Literature Review

"When you want to know how things really work, study them when they're coming apart."

– William Gibson

2.1 Introduction

THIS chapter discusses the literature and findings related to this thesis and its applications. There are mainly three concepts provided in this chapter, aimed at a novice audience who will be able to follow along with the ideas behind this thesis. First, a brief description of the initial developments of microscopes, their functionality in laboratories, and methods for their successful usage are provided. Secondly, the modernization of microscopes and their worldwide adoption by the pathology community as well as researchers is discussed. Finally, the results of some breakthrough research in histopathology using advanced algorithms are presented; the opportunities that AI brings to the detached areas of pathology are discussed. The area between pathologists and engineers in research is an exciting area with ample room for the advancement of the certainty of integration of digital systems into pathology.

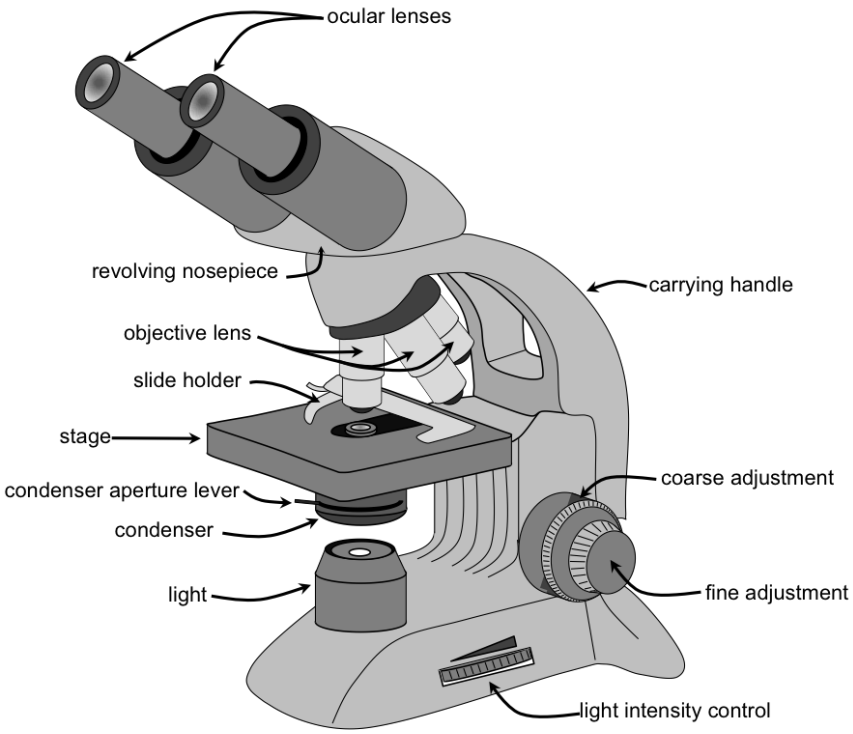


Figure 2.1: Complete anatomy of an optical light microscope. Image adapted from: http://thebiologyprimer.com/microscopy-lab.

2.2 Light Microscopy in Pathology

Remarkable developments in and modernization of microscopes was achieved in the 20th century, broadening the scope and accessibility of unanswered questions regarding the causes of disease and illness. The concept of studying tissue cells under a microscope is defined as "microscopy". Robert Hooke's "Micrographia" is considered to be the scientific backbone of microscopy [20]. Since its inception, various types of microscopes and lenses have been developed to increase the practical understanding of disease theories. An existing application of microscopy is the domain integration of CAD to automate laborious tasks. Basic light microscopes are prevalent in the majority of hospitals and laboratories because of their affordability and simplicity [21]. A light optical microscope and its auxiliary components are shown in Figure 2.1. [1]

[1] https://en.wikipedia.org/wiki/Timeline_of_microscope_technology

2.2.1 Working Principles of Microscopes

Optical microscopes work on the principle of light refraction. The phenomenon of interaction of light rays with a specimen is captured with a series of optical lenses, which allows a magnified view of the specimen. Two parameters that are important in microscopy are **magnification** and **resolution**.

Magnification is measured by how large a microscope can cause an object to appear. Light microscopes typically used in pathology laboratories magnify up to approximately 1000 times the actual size, enlarging a specimen of 1 mm × 1 mm to 1000 mm × 1000 mm in the microscope image [2]. In traditional microscopes, the magnification of the objective lens depends on the individual magnifications of the objective lens and eyepiece lens. Equation 2.1 gives the total magnification of an eyepiece lens [22].

$$\text{Total Magnification} = \frac{d}{f_o} \times M_e, \tag{2.1}$$

where d is the distance between the objective back focal plane and focal plane of the eyepiece, f_o is the focal length of the objective lens, and M_e is the fixed magnification of the eyepiece.

Resolution of a microscope or lens is the smallest distance by which two points can be separated and still be distinguished as separate objects. The smaller this value, the higher the resolving power of the microscope and the better the clarity and detail of the image. Two cell nuclei very close together on a specimen slide might look like a single blurry dot through a microscope with low resolving power but could be easily distinguished as individual cell nuclei on a microscope with high resolving power [22].

2.2.2 Glass Slide Preparation

The process of glass slide preparation, before tissue specimens can be clinically examined, can be divided into four parts [23]:

1. Collection: Excision of tissue specimen in situ by a surgeon, either by dissection or needle biopsy.

[2] The actual magnification in digital monitors is influenced by the dots per inch (DPI) or pixels per inch, which can alter the actual viewing magnifications in digital mode

2. Embedding: The tissue sample is settled in a mixture of specifically designed chemicals to react with it. This reaction enhances the visual colors of the tissue along with its retained characteristics at the molecular level, allowing fine-grained diagnosis.

3. Section: The samples embedded are often cut into thin slices of 5 to 10 mm, and each tissue preparation is called a "glass slide".

4. Staining : A variety of chemicals and dyes are applied to glass slides to distinguish the tissue contents. Hematoxylin and eosin (H&E) is the most commonly used stain to enhance the colors of the nucleus as purple and extracellular matrix and cytoplasm as pink[3]. This allows for better semantic differentiation to understand the morphology.

Finally, the prepared slide is inspected by a domain expert pathologist, and the final findings are rendered in the pathology report.

2.2.3 Transition to Digital Pathology

In this section, the major transformation from traditional to digital microscopy is discussed. The impact of digitization in clinical pathology is not only limited to diagnosis but also benefits therapy settings, tele-medicine, and case consultation, and has driven huge interest in research on the validation of complete virtual microscopy.

Photomicroscopy

Traditional microscopy has persisted along with the integration of digital systems. Microscopy has emerged with high-resolution optics to obtain more enhanced, clear, and unambiguous digital images. Digitization in microscopy began with the ability to attach digital cameras to a microscope. This enabled pathologists to capture specific regions observed from the eyepiece. Irrespective of standardized brands or technologies, any combination of a digital camera and optical microscope can be used to secure a digital image [24].

Specialized microscopic camera units are available for capturing images, as these systems allow stabilized images with auto-focus functionality. Common cameras used for capturing photographs can be economical, but only at the cost of lower image quality. A comparative analysis between these two types of image capturing systems is provided in [25], which promotes investment in better digital infrastructure that yields better image support and diagnosis [26].

[3] https://serc.carleton.edu/18523

With the increasing popularity of smartphones and high-definition mobile cameras, a microscopic system based on mobile cameras can meet the compelling demand for disease diagnosis with limited resources [27, 28]. At least a 5 MP camera is required for close diffraction-limited resolution images over a wide range of magnifications, including those relevant for single-cell imaging [29]. The resultant image can be a few megabytes in a standard JPEG or PNG format and can be accessed without any special software. Despite being small in memory size, these images can be easily transmitted by e-mail or a smartphone as an attachment for further examination by an expert at a remote location.

WSI Scanners

WSI, also known as virtual microscopy or digital microscopy, is an emerging technology used for converting a glass slide into a virtual slide, commonly known as a whole-slide image. These digital slides are used by experts to conduct observations using specialized software or for image analysis. The modality responsible for preparing the image is called a WSI scanner, and it consists of robotic microscopes that scan a glass slide. A composite whole-slide image is obtained by scanning small areas of glass slides individually and merging them using sophisticated software [30].

Technological advancements in robotics and control systems have significantly advanced WSI technology. These improvements have reduced the time required for glass slide scanning (average time has reduced to a few minutes per slide) and the automatic processing of more than 300 glass slides [31, 32, 33].

Portable whole-slide scanners can be set up easily on normal tables without occupying a large space in diagnostic centers. These modern WSI scanners convert glass slides in efficient time, automatically with no human intervention for tissue localization or focus plane selection [34].

Virtual slides obtained from WSI scanners consume more memory than microscopic snapshots [30]. WSI images can be 2-4 GB with a resolution of 50,000 x 50,000 pixels. WSI images are saved as specially-designed format files with the capability of saving multiple tiles in one file. The arrangement of the image tiles resembles the shape of a pyramid [35].

Current State of Microscopy

This section contains a brief literature survey of current microscopy research. The advantages of advanced digital integration in the workflow and their consequences are discussed.

According to the literature, there are several factors that affect the adoption of virtual microscopy. A major study [36] focused on the time variation between diagnosing a glass slide and a virtual slide for skin cancer. The results were highly biased as pathologists were easily able to reach a diagnosis on glass slides in an average time of 137.61 s, while for a virtual slide, the time increased to 176.77 s. This increased time acts as a barrier for the adoption of virtual microscopy. Various factors, such as increased WSI system setup cost, limited control for pathologists while viewing virtual slides, ambiguous quality of WSI slides compared to glass slides, and inability of pathologists to operate specialized WSI system software, limit the wide acceptability of virtual microscopy across small- and medium-scale laboratories [37].

Recent studies have compared virtual microscopy to traditional microscopy. Non-inferior results were obtained for virtual microscopy in several studies [38, 39, 40]. The positive results support the transition of the pathology environment from traditional microscopy to virtual microscopy, to help meet the compelling demand.

2.2.4 Effects of Magnification in Pathology

Magnification is a transformation in the actual size of an object, performed so that the object can be visibly enlarged. In traditional light microscopes used for pathology, a wide range of lens powers are available to ensure that the reflection of tissue observed through the oculars is large enough for clear examination. In contrast to LM, for virtual microscopy systems, magnification means matching the digital and optical resolution of the scanning camera and digital monitor screen. Once the image has been scanned with a high-quality sensor, additional magnification becomes redundant. Images on a digital display can turn out to be very different from the microscopic view when the modality to observe the specimen is altered from traditional light microscopy to recently introduced high-end WSI technology. The resulting magnification power varies based on image modalities; for virtual microscopy, the final magnification level is proportional to the distance between the screen and viewer, whereas in traditional microscopy, the resultant magnification power is derived from Equation 2.1. This variation stops magnification power from being a common unit across all modalities. In whole-slide scanners, vendor-specific micron pixel sensors lead to variations in the resultant magnification. As a result, a new standardized unit that can generalize magnification level from both modalities must be accepted universally to make diagnosis and research more ubiquitous [41].

The transition from traditional microscopy to digital microscopy is due to the availability of powerful computers that help pathologists to inspect and analyze specimens on

digital screens, eliminating the need for microscopy and glass slides. Although some medical institutions are adopting complete digital microscope systems in place of traditional microscopy, there may be a risk of stand-alone technology. As computer algorithms are designed to process static images, they fail to learn the semantic chronology possessed by moving the tissue and switching the magnification of a microscope. Static images on computer monitors certainly augment the laboratory environment. Few pathologist believe that studying slides of human tissue under a microscope adds a unique experience that is not offered by static images [42].

2.2.5 Advances in Digital Pathology

Microscopy is the primary source of almost all types of disease diagnosis, including all types of cancer [43]. With an increasing workload in the existing work environment, automating microscopy is essential. Integrating automation systems simply involves optimizing existing methods in clinical settings, which can improve existing systems.

Significant improvement in diagnostic accuracy is achieved by implementing strategies such as case conferences, consultations, and secondary diagnosis. The authors in [44] conducted error analysis by implementing such strategies and obtained errors in the range of 1.2 to 50 per 1000 patients. Reduced error rates will assist in the smooth acceptance of ML techniques in clinical settings, facilitating a better future for histopathology.

Applications of ML in Histopathology

This section examines the recent methods proposed for image analysis. A brief discussion of ML techniques used in DP images for feature extraction, image similarity matching systems, segmentation, and tissue level diagnosis is provided.

The initial idea of digital image analysis was not for the task of natural scenic or objective images but for developing automated image analysis of ultrasound medical images [45]. A comprehensive survey by Gurcan et al. [7] lists the applications of CAD systems deployed in healthcare since the 1990s. In the last decade, there has been exponential integration of CAD systems in radiology [46]. With increased efficiency achieved in radiology, pathology has improved as well.

With increasing workload in pathology routines, there is an evident need to meet demands by integrating smart computers with a pathology routine to automate laborious tasks, allowing pathologists to focus more on complex cases[47, 48]. Increased acceptance of WSI systems and their applications in pathology has accumulated substantial medical

images, increasing the pressure for detailed analysis and improving the current state of diagnosis [45, 49, 50]. The benefits of quantitative analysis of pathology imaging using ML techniques are being acknowledged by researchers from academia as well as industry [7].

Image Analysis in Digital Pathology

The main gateways to advancing ML applications in DP are the image dimensions and extensive requirements of computational resources. The primary step for any histopathological diagnosis is to learn the visual semantic features at the cellular level. Particularly, cell structures, nuclei size and shape, cellular matrix, glands, and lymphocytes are prominent features that can be used for cancer diagnosis. Owing to the limited computational resources and technical complexity faced by virtual microscopy, the idea of integrating CAD systems for automated cancer diagnosis is awaiting necessary validation for successful deployment in laboratories [51, 52]. With increased accessibility of computational power, recent literature has provided astonishing results for segmenting nuclei from tissue images [53]. Automated tissue segmentation using CNNs with the U-Net architecture was developed for skin cancer detection. The proposed technique outperformed existing techniques [54]. Detecting the region of interest (ROI), the possibly tumor-affected region, in WSI using CNNs trained on non-histopathology images [55] and gland segmentation[56, 57] are recent applications of ML in task-specific DP.

Shivang et al. [58] demonstrated automated cancer grading and localization techniques for prostate and breast glands. They also tried to recognize the presence of malignancy, and favorably consistent results were obtained compared to manual segmentation. Another study in [59], compared the efficiency of AI for the Gleason grading of prostate slides, with 0.96 correlation between the system and an expert pathologist. A reduction in diagnostic error of up to 85% was observed when Wang et al. proposed DL-based CAD system for identifying metastatic breast cancer[60]. A breakthrough in gigapixel image processing was achieved by Liu et al. [61], localizing 100×100 pixels in a virtual slide. Their results were achieved using the CNN architecture.

A rotational invariant CNN model was developed in [62] for lymph node tumor detection. An evident increase in accuracy was observed with the use of rotation-invariant features. A comparative study of feature extraction techniques to obtain morphological features from images to detect metastasis was conducted by Kumar et al. [63]. Comparative studies on feature extraction techniques for medical tasks have been conducted by several researchers [47, 64].

The influence of magnification level on breast cancer diagnosis was evaluated by Bayramoglu et al. [65] and Gupta et al. [66]. An end-to-end multitasking CNN-based classifi-

cation system for magnification analysis along with malignancy detection was presented by [65]. Otlara et al. focused only on regressing the magnification level, implementing a CNN-based regression technique by fine-tuning bottleneck layers in a deep CNN.

Content-Based Image Retrieval Retrieving similar images based on image morphology requires the characterization of semantic information that is specific to an image. CBIR needs access to huge data repositories with high-quality images [67, 14]. Most research on WSI technology is done on image patches which are subset of one gigapixel WSI image.

An online CBIR system was developed by Zheng et al., wherein a user would upload a query image and its corresponding search parameters. The algorithm provides similar images based on the color histogram, texture, and wavelet coefficients [68]. A multi-tiered CBIR system was developed for virtual slide processing, which is able to classify and retrieve similar image slides of a whole slide or multiple image patches [69]. Unsupervised feature learning combined with the bag-of-features concept was implemented for feature matching in CBIR, competing with existing state-of-the-art canonical patch descriptors [70].

The following sections describe the well-established approaches used for extraction of semantic knowledge infused in images. In the image processing domain, this process is known as feature extraction, which is responsible for transforming raw data using a transformation function. This function is directly responsible for dimensionality reduction of raw images to a 1-dimensional feature vector. This feature vector is supposed to encompass the image-specific shapes, edges, and typical morphological structures.

2.3 Image Representation

2.3.1 Handcrafted Features

This section discusses the characterization of images using the traditional algorithms used in this thesis. *Hand-Crafted Features* are features obtained by applying a series of mathematical functions across pixel values of an image. These algorithms were first manually developed for their target tasks, while their application to universal tasks was developed later. Over the last few decades, handcrafted features have been the primary source of image representation, and their popularity has extended their application, even after the progress of deep embeddings from DL. It is a common approach in image analysis research to compare traditional computer vision algorithms with deep descriptors for various machine learning tasks [63, 64, 71].

Local Binary Patterns (LBP)

LBP is a simple feature extraction technique with an embedded mathematical function familiarized by Ojala et al. in 2002 [72]. Because of its static function, it does not require any training for its application. As a result, its features are categorized as hand crafted.

Various versions of LBP algorithms have been developed, such as uniform and non-uniform LBP. These variants significantly increase the scope of applications[73, 74, 75]. Owing to its simplicity, good performance, and competitive results, this thesis implements uniform LBP as a feature extractor for selective experiments.

LBP is responsible for local instead of global representation of images. The LBP function transforms each pixel value from an image with a new intensity value calculated based on its neighboring pixel values from the same image.

A uniform LBP is defined by the transitions of 0 to 1 or vice versa when the LBP operator is traversed circularly. For example, the patterns 00000000 (0 transitions) and 00011100 (2 transitions) are uniform, whereas the patterns 10011001 (4 transitions) and 00101010 (6 transitions) are non-uniform. For application of the uniform LBP algorithm, grayscale images are required for this algorithm to perform well. The binary patterns are obtained by applying a circular operator to every pixel value of the image in the first step. An aggregation of all non-uniform LBP patterns from the binary patterns is considered as the final histogram.

LBP Operator: In the uniform LBP algorithm, a circular LBP operator is used for the selection of neighborhood pixels. The circular operator is denoted by $LBP_{P,R}$, where P explicitly defines the finite number of neighboring pixels and R controls the radius of the circular operator to locate the pixel information. The number of neighbors P and radius R are two parameters that control the selection of neighboring pixels. In Figure 2.2, an image is represented by a grid, where each square is an individual pixel from the image, and the circular operator with center pixel g_c is represented in black, while neighboring pixels, represented by $g_0, g_1, ..., g_{P-1}$, are colored in green. The parameter settings of the three combinations are shown in Figure 2.2. The neighbors are equidistant across the circumference of the imaginary circle, and the coordinates of each i^{th} pixel on the circle C are represented by g_t and can be derived by

$$g_t = \left(-r \sin\left(\frac{2\pi i}{P}\right), r \cos\left(\frac{2\pi i}{P}\right)\right). \tag{2.2}$$

For pixel coordinates that are located on the boundaries of the grid, their values are

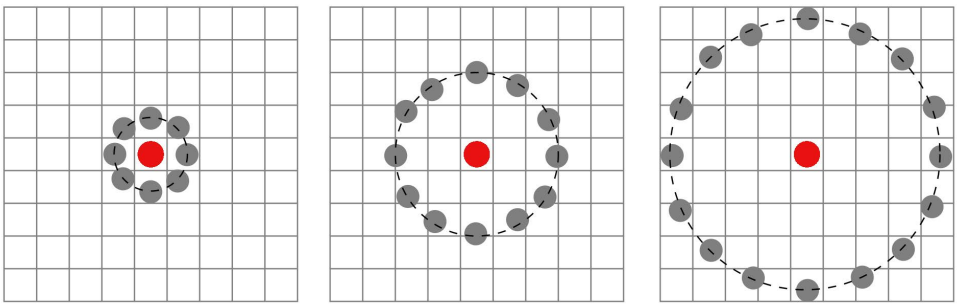


Figure 2.2: LBP: circular operator with three combinations of the radius (r) and neighbor (n) parameters. **Left**: operator with radius $r = 1$ and neighbor $n = 8$. **Middle**: operator with radius $r = 2$ and neighbor $n = 16$. **Right**: operator with radius $r = 3$ and neighbor $n = 24$. Neighbor pixels are colored gray, while the center pixel is colored in red.

interpolated. The pixel values obtained using Equation 2.2 surpass the circular operator function to calculate the binary transitions. The length of the array of binary transitions is $P-1$, and Equation 2.3 derives the mathematical function to obtain a binary transition for each pixel.

$$LBP_{P,R} = \sum_{P=0}^{P-1} f\left(g_P - g_c\right) 2^P, \tag{2.3}$$

where

$$f(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The function described in Equation 2.3 compares the intensity of neighboring pixels with that of the center pixel. For a pixel value greater than or equal to that of the center pixel g_c, the corresponding bit in the binary mask is set to 1, and the binary mask is set to 0. Binary transitions for P bits are obtained across all g_{P-1} pixels and further aggregated into histograms, which are considered as the final features.

LBP Histogram: The binary transitions are P bit binary arrays that are converted to rational numbers in the range $[0, 2^p]$. The absolute number of binary transitions is denoted by $\eta_{\mu,P}$, where μ represents the uniform LBP scheme, and Equation 2.4 gives the

#	LBP Parameters		Feature Length
	Radius (r)	Neighbors (n)	(l)
LBPγ_1	1	8	59
LBPγ_2	2	16	243
LBPγ_3	3	24	555

Table 2.1: Parameters of LBP feature extraction with resultant length of feature vector.

mathematical function to calculate total binary transitions.

$$\eta_{\mu,P} = P \times (P-1) + 3 \tag{2.4}$$

The mapping function U_p is responsible for mapping all the binary patterns with their associated labels in the look-up table. This look-up table contains 2^P elements. All the elements are assigned unique labels for each uniform pattern, and all the non-uniform labels are aggregated in one bin; the final histogram is the evident feature of the input image. Therefore, the total number of histogram bins in a uniform LBP is $\eta_{\mu,P} + 1$ (i.e., all individual bins for $\eta_{\mu,P}$ and 1 bin for all non-uniform masks).

The process of feature extraction using the uniform LBP algorithm can be summarized by the following steps:

1. Build the look-up table U_p for all P values.
2. For image I_m, apply circular operator $LBP_{P,R}$ for all $I_{x,y}$ and save the LBP masks in an array M_i.
3. Check M_i for uniformity using function U_p and store the results in another array h_i
4. Feature vector lbp_f of image I_m is the histogram of the aggregated array of unique values from array h_i.

The length of the feature vector deviates significantly when the parameters P and R (sampling of neighborhood pixels and radius of circular LBP operator, respectively) are changed. The uniform LBP algorithm has been successful in many classification tasks, and the values of (8,1) for parameters (P, R) could capture 90% of image characteristics in the work of Ojala et al. [72]. The LBP algorithm is fast in computation with a time complexity of $O(g)$, where g is the number of pixels, and has a simplistic function-based approach, which make it an optimal choice of algorithm for this thesis.

2.3.2 Deep Embeddings

With increasing computational resources in terms of processing power (i. e., accessibility of cloud computing and GPU server availability) and memory storage in the terabytes, the applications of deep CNN have become easy, which has decreased the popularity of handcrafted features. The peak performance of CNNs relies on domain-specific training, enabling the network to learn information from domain-specific images. Transfer learning and fine-tuning are other approaches to CNN implementation with limited data points. The following section discusses the DL approaches and CNN fundamentals required for implementing them as feature extractors.

CNN Fundamentals

The CNN model is built by stacking various layers with forward connections. The input layer of the CNN model accepts a 3^{rd} order tensor of image height, image width, and channels. The image tensors are iterated through the layers of the model, and the final predictions are obtained as output. Forward pass (FP), error function, and backward pass (BP) are three fundamental concepts of neural networks and are defined as follows:

1. FP refers to the calculation process of output values based on the input data. The final output value is calculated by traversing all layers from first to last, forming an open loop structure. The output values $\hat{p}$ are the sum of all weighted inputs transformed across a nonlinear function. This nonlinear function can be a sigmoid function, tanh, or softmax depending on the problem in question. The mathematical derivation of the predicted outputs obtained using only FP is given as:

$$x^1 \rightarrow w^1 \rightarrow x^2 ... x^N \rightarrow w^N \rightarrow \hat{p}, \tag{2.5}$$

   where $f(\cdot)$ is the activation function, x^i is the input to the i^{th} layer, w^i is the weight of the i^{th} layer, and $\hat{p}$ is the calculated output of the CNN.

2. Error function refers to calculation of model loss: the difference between actual output and predicted output. The error function or loss function is defined by:

$$\gamma(w) = p - \hat{p}. \tag{2.6}$$

   The main purpose of a loss function is to minimize the error values between the actual values p and predicted values $\hat{p}$. There are several loss functions targeted at a

specific set of problems. For instance, the "binary crossentropy" loss function is used specifically for binary classification problems.

3. Backward Pass (BP) refers to the process of adjusting model parameters(i.e., weights) using an optimization function(e.g., stochastic gradient descent). The error is iterated through the previous layers in the reverse direction, where it adjusts the connecting weights and bias resulting in a minimized error.

$$w^1 \xleftarrow{\frac{\partial \gamma}{\partial x^2}} w^2 \xleftarrow{\frac{\partial \gamma}{\partial x^3}} ... \xleftarrow{\frac{\partial \gamma}{\partial x^N}} w^N \xleftarrow{\frac{\partial \gamma}{\partial p}} . \tag{2.7}$$

   To minimize the loss, gradient descent optimization iteratively reduces the discrepancy by moving in the direction opposite to the steepest slope. The loss at each iteration is calculated using the partial derivative of the loss function with respect to the weights. These iterations are repeated to adjust the weights until the global minimum error is achieved. The global minimum error provides the best-fit model for the defined problem.

A common approach in the DL community is the interchangeability or reusability of model parameters in contrasting domains. This approach has applied CNNs to feature extraction, transfer learning, and fine-tuning. The following section summarizes the implementation of a CNN for extracting features and application as a classification model.

Feature Extraction

The architecture of neural networks is highly modular. The quality of information obtained from input images increases successively owing to the multi-layer topology. CNN features are invariant to translation due to the presence of convolutional layers. A CNN provides the flexibility to characterize an image using activation values from any layer of the model. Low-level information (i.e., edges and corners) is captured in the initial layers, while high-level information (i.e., tissue texture and image-specific pattern) is captured in the final layers. Most feature extraction applications seem to use features from the last layer of the neural network model.

In this thesis, 4 models of pre-trained neural networks are investigated to classify the magnification level of histopathology images. The features from these models are referred to as deep embeddings. Table 2.2 yields information regarding the layer number, whose activation values are calculated as deep features for each model.

Feature Model	Layer (#)	Descriptor Length
DenseNet	121	1024
ResNet	151	2048
EfficientNet	237	1280
VGG	19	512

Table 2.2: List of pre-trained models used for extracting features. Additional information includes the layer number (#) used for calculating features and resulting length of the $1-d$ feature vector for all images.

Models The models used for feature extraction are VGG19, DenseNet121, ResNet151, and EfficientNet. These models are trained using ImageNet data of natural classes and used for image representation of histopathology images.

Implementation The images from all datasets described in Chapter 3 are traversed through all models using an FP to represent the input images in the form of activation values from the last pooling layer of each model. Table 2.2 describes the layer used for extracting features and its length of features from each model.

Fine-tuning Fine-tuning is a technique that uses discrete layers of a model with pre-trained weights, while the rest of the model is trained based on the domain-specific data. In this way, the model adapts to domain knowledge by adjusting the weights of the trainable layers (i.e., layers except those with ImageNet weights), allowing room for increased model performance. In this thesis, VGG19 and DenseNet121 are fine-tuned for the adaption of histopathology images. VGG19 is fine-tuned from layer 16 onwards, and the last block of DenseNet is re-trained for building consensus using DenseNet121. Bottleneck layers are the fully connected layers responsible for calculating the predictions. The configuration of fully connected layers is identical across all fine-tuning experiments in this thesis. A common practice seen in model fine-tuning of a CNN is adding a normalization and dropout layer between densely connected layers. This allows the network to learn better discriminative features, giving accurate predictions on unseen images. Table 2.3 describes the layers used with additional parameters added to the network. The input layer is the first dense layer, with neurons equal to the total length of the feature vector from the model, and is connected to a second dense layer with 512 neurons. The values from the previous layer are batch normalized to contain the length of the features, and 50% arbitrary values are dropped out to avoid over-fitting. Left activation values are passed to a dense layer of 256 neurons, which are directly connected to the output layer with 4 neurons activated by the softmax

Layer (type)	Output Shape	Param (#)
VGG19 (Model Input Layer)	(None, 512)	20024384
Dense Layer	(None, 512)	262656
Batch Normalization	(None, 512)	2048
Dropout (0.5)	(None, 512)	0
Dense Layer	(None, 256)	131328
Dense Layer (Output)	(None, 4)	1028
Trainable params:		396,036

Table 2.3: Bottleneck configuration for fine-tuning VGG19 model. Total additional parameters due to newly added layers are provided under "Trainable params". The output shape of each layer contains the **None** value, which is a variable based on the batch size during experiments

function to predict the class labels in terms of a probability distribution. All dense layers except the output layer are activated using the ReLU activation function.

2.4 Image Classification

The area of machine learning that learns a function that maps an input (i.e., feature vector) to an output (i.e., labels) based on an example input—output pair is called supervised learning, and algorithms implementing supervised learning are called classification algorithms [76]. A classification algorithm analyzes the training data and approximates a function that can be used for mapping new test samples. To predict the correct label for unseen samples, the learning algorithm must be generalized from the training data. The features obtained from the discussed feature extraction techniques are considered as inputs for classifier training. Classifier performance is highly affected by its internal parameters, and each classification algorithm has its specific set of criteria responsible for adjusting the parameter values. The classifiers used for conducting experiments in this thesis are described in the following sections.

2.4.1 k-Nearest Neighbors (k-NN)

The k-NN classifier [77] is instance-based or non-generalizing, where the function is only estimated locally and the computation is deferred until function evaluation is called for an unseen data sample. An unseen sample point is mapped in the feature space, and the class

assignment is computed by the majority vote of the nearest neighbors. The number of neighbors (k) and the distance metric for neighbor assignment to be considered for voting are specified explicitly by the user; a large k value suppresses the noise, making inter-class boundaries less distinct.

2.4.2 Random Forest (RF)

RF classifiers [78] are based on the concept of multiple decision trees merged for classifying the problem space. A finite number (n) of individual trees are built, and the majority vote across all individual trees accounts for the final class assignment. RF classifiers add randomness to the model by randomly sub-sampling the feature space while growing trees. The trees are grown by providing a feature–label pair to the root node, which splits the feature space based on the loss function. These splits are known as leaf nodes, with each leaf node accumulating a sub-space of features. This chain continues until a stopping criterion is met. For any query image, features are traversed down the tree, and the label from the traveled leaf node is assigned to the query image. Important parameters for RF classifiers are the number of estimators, information gain criterion, and maximum depth to which a tree can grow.

2.4.3 Support Vector Machine (SVM)

In a SVM [79], each feature vector is viewed as a data point (p) from a feature–label subspace, and the classifier is responsible for constructing a single $(p-1)$-dimensional hyperplane, which can separate all classes of data points. There are multiple hyperplanes that might classify the data point in a high-dimensional feature space. One reasonable choice for the best hyperplane is the one that represents the largest separation or margin between the two classes. Hence, the hyperplane with the maximum distance from its boundary to the nearest data point on each side is selected. Kernel functions are responsible for achieving a nonlinear hyperplane when classifying non-binary problems. A few important parameters that strongly affect SVM performance are regularization parameter limiting the width of the boundary and kernel function, which are responsible for nonlinear class boundaries.

2.5 Final Remarks

AI algorithms that leverage the biological concept of neurons in designing algorithms are known as Deep Learning. In a comprehensive review by Litjens et al. [80], it is evident that DL techniques have huge potential for applications in medical imaging. The increase in publications on medical image analysis since 2015 proves the evolution of pathology. Various conferences have special issues specifically targeted at medical images, such as IEEE Transactions on Medical Imaging [81].

Contradicting all the achievements of ML, lack of generalization is seen because of the highly sensitive nature of images; tiny alterations to input images can cause existing state-of-the-art algorithms to fail [82].

In this work, multiple techniques are studied to recognize the magnification level of histopathology images. An independent study to analyze the influence of the primary site of an image in recognizing magnification level is conducted extensively. In the next chapter, the prevalence and shortcomings of popular datasets used for medical imaging research are explained and two datasets are proposed.

Chapter 3

Datasets and Evaluation Criteria

"Failure is the opportunity to begin again more intelligently."

– Henry Ford

3.1 Introduction

THIS chapter presents the datasets used in the experiments conducted in this study. A brief description and some limitations of publicly available datasets are presented, and the detailed procedure followed in designing the proposed dataset is discussed. The development of the KIMIA-MAG-5 and OMAX data repositories was conducted at Kimia Lab, University of Waterloo.

The *KIMIA-MAG-5* dataset is composed of 33,345 manually selected image patches.[1] The repository is 5 GB and will be publicly available for education and research upon request. A dedicated test dataset for evaluating model performance was maintained independently from the images used for training.

3.2 Motivation

All publicly available data repositories are specifically designed for a primary goal. Currently, the Cancer Genome Atlas (TCGA) [3] is the largest public repository, with more

[1]The paper presenting KIMIA-MAG-5 is under review for publication in ICTAI 2020

than $10k$ whole-slide scans. The major limitation of these images is the lack of annotation, as the slides do not show the actual Region of Interest (ROI), which is important for diagnosis. The fundamental limitations of *BreakHis* images are low resolution and a lack of diversified organs. To mitigate these limitations, there is a pressing need for a new dataset. KIMIA-MAG-5 is the proposed solution for understanding the practical effects of 25 distinct organs for magnification prediction. Particular focus was dedicated to including maximum variability in terms of color and texture patterns persistent across all primary sites. Manual selection of patches adds exquisite value to the dataset with the least chance of marker or tissue fold presence.

3.3 KIMIA-MAG-5 Dataset

When designing the KIMIA-MAG-5 dataset, a well-planned procedure was followed to generate a diverse balanced dataset to evaluate the magnification scale and conduct origin of tissue (i.e., organ) analysis for the task of magnification prediction. Virtual slides from TCGA were collected for 25 organs (or primary sites in clinical terminology) to develop this dataset [3].

KIMIA-MAG-5 is a subset of this large data repository. Manual filtering of patches for removal of tissue markers, low tissue-background ratio, and tissue folds was conducted to prepare a clean, high-quality data repository for research. The main motivation for investing time and effort in the design of this dataset was to improve the shortage of microscopic-style multi-magnification images. This approach replicates to the nearest level to enable a comparison between traditional microscopy and virtual microscopy.

The complete design procedure followed for developing KIMIA-MAG-5 is as follows:

1. Coordinates X, Y were randomly selected for each slide. Regions of 800 $\times$ 800 pixels were read from each slide at either 40x or 20x magnification.

2. The coordinates were traversed along the axis to lower magnifications of 20x, 10x, 5x, and 2.5x to read the area from the same image center. For each magnification level, three patches were extracted, resulting in a total of 34,175 image patches. All images at each magnification were saved as JPEG files.

3. Minimum tissue-background ratio for all images was set at 51%. This ratio allows inclusion of some background areas in low-magnification images. An algorithm was deployed for filtering images with a low tissue-background ratio.

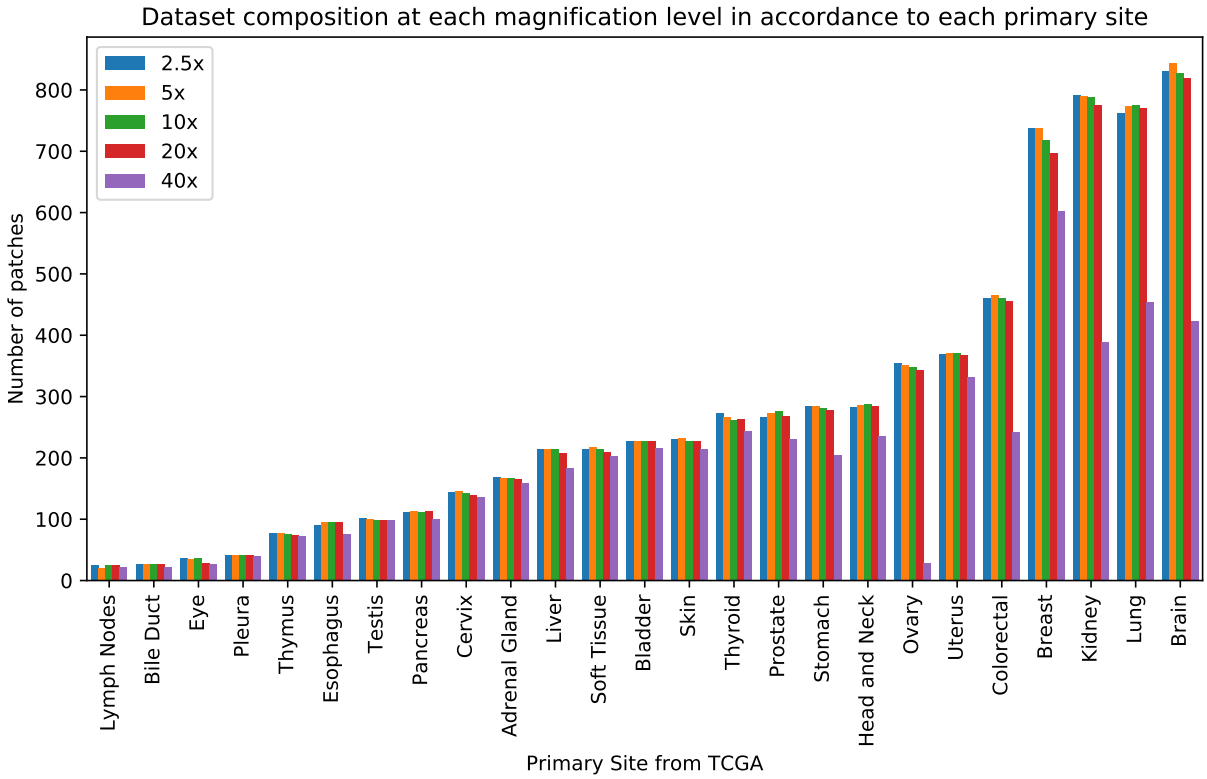


Figure 3.1: Histogram with 25 bins depicting image distribution for all primary sites with stacked bars for magnification levels present in the KIMIA-MAG-5 dataset. The X-axis indicates the primary sites, and the Y-axis denotes the number of images for each site

4. While patching, some images with markers or sketch-pen marks, artefacts, blurred regions, and tissue folds were captured. Manual removal of such unwanted images was conducted by visually checking the entire dataset.

5. In total, 33,345 patches from 2,862 patients' slides from 25 primary sites were obtained at five unique magnifications.

6. 20% of patient images were kept independent of training to evaluate model performance.

In total, KIMIA-MAG-5 contains 7,214 patches at 2.5x magnification, 7,157 patches at 5x, 7,100 patches at 10x, 7,008 patches at 20x, and 4,956 patches at 40x. The statistics of the primary site distribution at all magnification levels are illustrated in Figure 3.1.

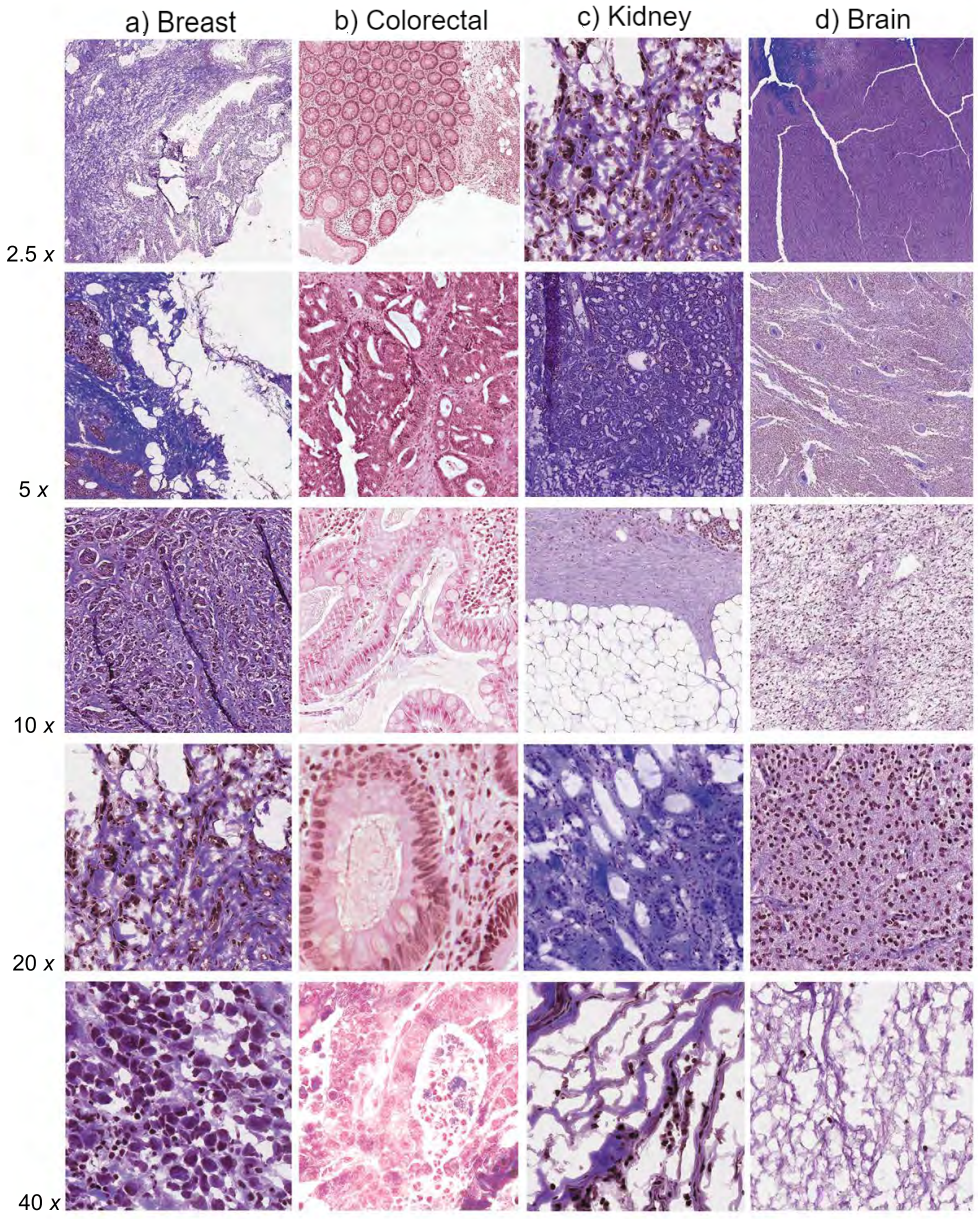


Figure 3.2: Image samples from four primary sites from KIMIA-MAG-5 dataset. Four columns represent a) Breast, b) Colorectal, c) Kidney, and d) Brain images. From left to right, the rows represent magnification levels of 2.5x, 5x 10x, 20x, and 40x.

Figure 3.2 shows 20 image samples from four primary sites: breast, colorectal, kidney, and brain. Color variation and increasing size of structural patterns are easily visible as one mavigates from lower to higher magnification.

3.4 BreakHis Dataset

The BreakHis dataset is a perfect example of magnification-influenced breast cancer classification using pathology images. The BreakHis dataset consists of images from 82 breast cancer patients. Images of these glass slides were captured at four magnification levels using a Samsung camera attached to a light microscope. Along with distinct magnification, these images are categorized as benign or malignant by a pathologist, while annotations of the actual tumor sites are missing, but a label for each image exists. The BreakHis dataset was developed by Spahnol et al., and a detailed description of classification experiments is presented in [83]. For the BreakHis dataset, patches from original images were cropped from the center to obtain square images of 460 × 460 pixels. In total, the BreakHis dataset consists of 7909 images with 1995 at 4x, 2081 at 10x, 2013 at 20x, and 1820 at 40x.

3.5 OMAX Dataset

The OMAX dataset is being designed and developed in-house at KIMIA Lab, University of Waterloo[2]. It is an evolving dataset, with a conscious effort made to represent the different textural patterns and types of tissues at multiple magnifications. The experiments conducted in this thesis employ images made available up to July $6^{th}, 2020$. In total, 64 glass slides were provided by Huron Digital Pathology[3]. The actual size of the tissue samples in real scale was 5 mm × 5 mm. The glass plate size was 7.5 cm × 2.5 cm, and tissue specimens were bounded in 1.5 cm × 1.5 cm square coverslips or circular coverslips with diameters of 1.5 cm. The primary motive for developing this dataset was to verify model capability by including diverse pathology stains captured by a microscope. The resulting dataset is heterogeneous in its color spectrum. Primary stains applied to glass slides used for capturing images were H&E, while some slides were stained with trichrome stains using Van Gieson's method. For images with H&E stains, purple and pink images were obtained, while yellow and brown images were obtained from trichrome stains. For capturing images, a Digital Compound Trinocular LED Lab Biological Microscope with

[2] https://kimialab.uwaterloo.ca/kimia/
[3] http://www.hurondigitalpathology.com/

a 14 MP camera and 1.25 NA lens from OMAX microscopes was used[4]. The following aspects were considered while capturing the images:

1. The patch coordinates x and y were randomly selected for each slide based on the tissue complexity.
2. For each slide, a total of $50 - 100$ snapshots were captured across all magnifications with some overlapping.
3. Snapshots captured using the camera-attached microscope had 4096×3288 pixels. For convenience, each image was cropped to 4000×3200 pixels, and 16 sub-images of size 1000×800 were taken from each snapshot.
4. In total, $17,952$ images were captured from 64 glass slides: 1696 at 4x magnification, 3504 at 10x, 5840 at 20x, and 6912 at 40X. The ratios of 10%, 15%, 30%, and 45% were preset for 4x,10x, 20x and 40x magnification, respectively.
5. 15% of primary sites were taken for testing; images from 10 glass slides were detached from the training dataset to evaluate the model performance.

Figure 3.3 shows image samples from each magnification level from 5 selected primary sites. Images with distinct visual patterns were manually selected for viewers to inspect the complexity in the patterns and color variations present in the dataset. Lymph node sections seen in column (d) of Figure 3.3 are yellowish in color due to different staining procedures applied to tissue specimens. These variations exhibit the heterogeneity possessed by OMAX dataset.

For the test split, 336 images at 4x magnification, 704 at 10x, 1264 at 20x, and 1024 at 40x, accounting for a total of 3328 images, were used for evaluating model performance. The contribution of images from each primary site is encompassed in the stacked histogram presented in Figure 3.4.

3.6 Evaluation Criteria

For any model, evaluation metrics are standardized units designed to analyze its performance. For the experiments conducted across all datasets, the proposed metric from Section 3.6.1, accuracy score, and F1-score from Sci-Kit were used to analyze the performance of the trained model [84].

[4] https://omaxmicroscope.com/

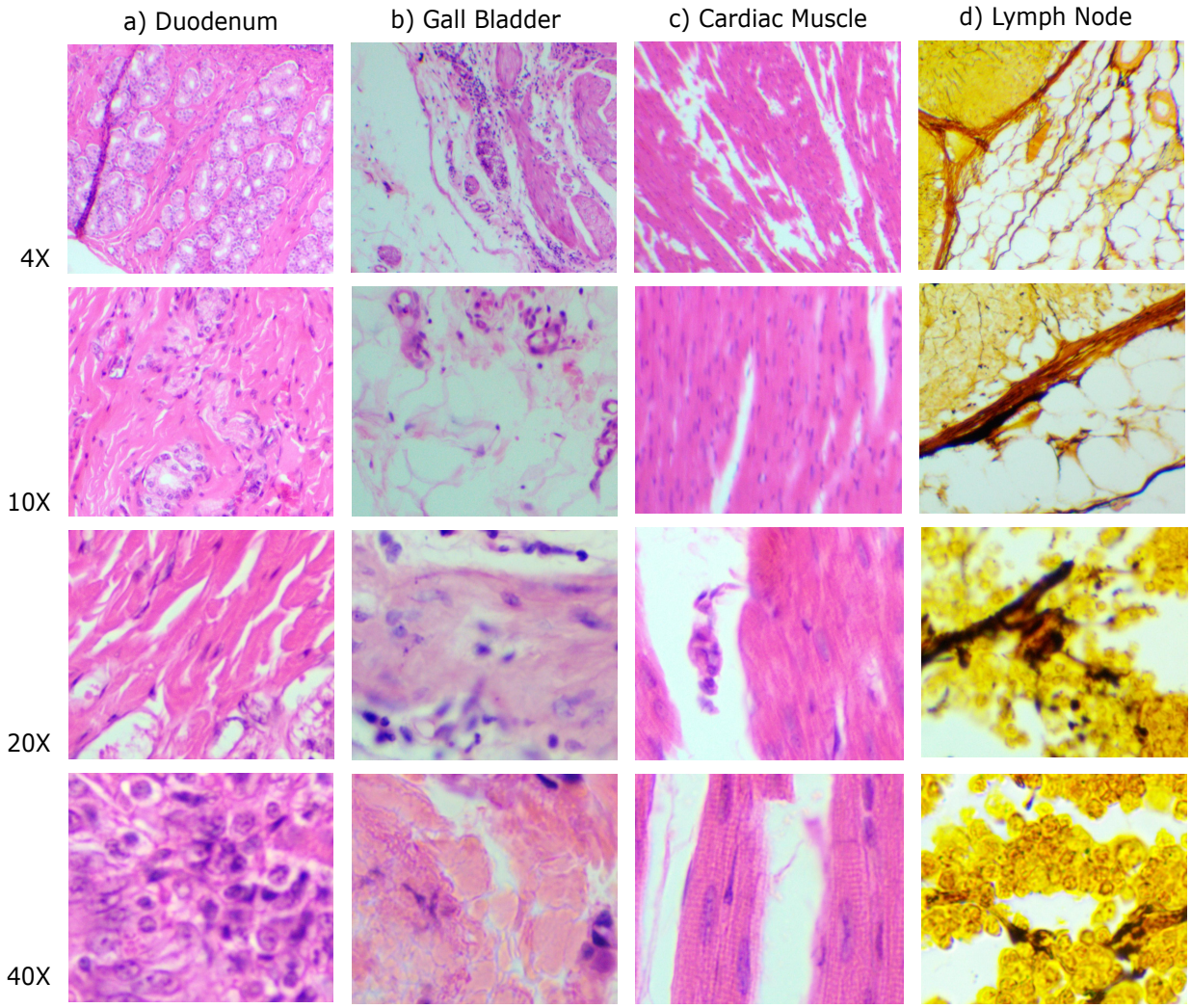


Figure 3.3: Image samples from OMAX dataset at four magnification levels of 4x, 10x, 20x, and 40x. Image samples from four primary sites are included as samples. For each primary site, the same coordinates are selected to compare visual characteristics with increasing magnification level. d) Lymph node sections are stained with trichrome dye, giving yellow-colored images in the OMAX dataset.

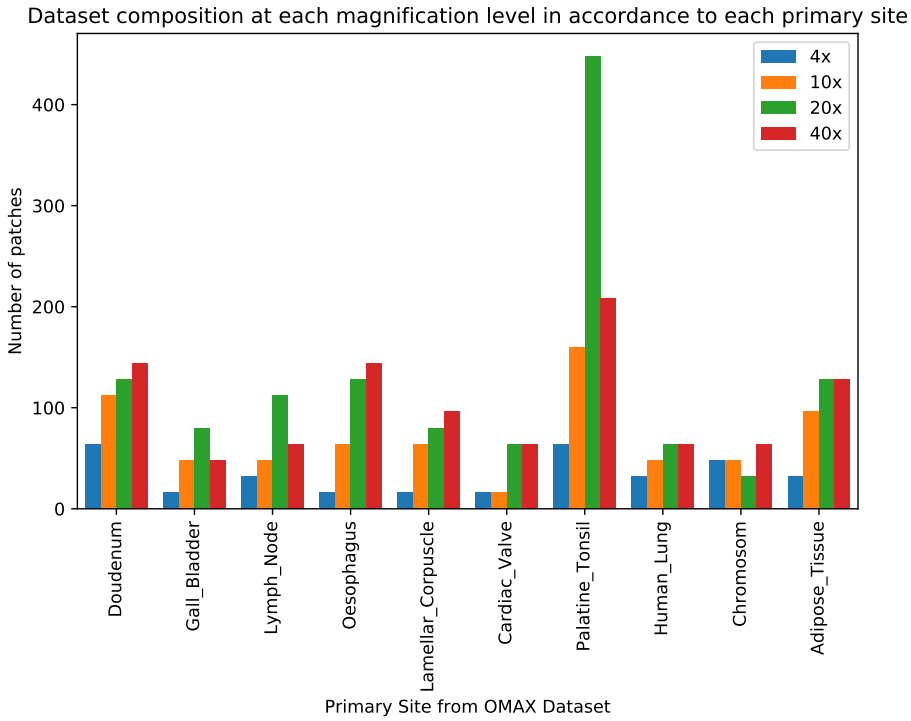


Figure 3.4: Histogram with 10 bins (primary sites) from OMAX data split used for model testing. Each bin is composed of four bar plots quantifying the number of patches associated with each magnification level for each primary site.

3.6.1 Accuracy

The primary analysis of misclassified magnification of test images of each glass slide was conducted using the proposed metric. In other words, the accuracy score at the individual patient level or primary site was calculated, which provides knowledge about the role of specific structures from specific primary sites responsible for erroneous results. Individual primary site analysis conducted using primary site accuracy P_{acc} is given in Equation 3.1. For magnification prediction analysis, total accuracy, represented by T_{acc}, was used. The calculation of T_{acc} can be obtained by replacing the term η_s from Equation 3.1 with η_{tot}. The mathematical derivations of both metrics, primary site accuracy P_{acc} and total accuracy T_{acc} are as follows:

$$P_{acc} = \sum_{s \epsilon S} \frac{(\tau_l^s \cap p_l^s)}{\eta_s} \tag{3.1}$$

$$T_{acc} = \frac{1}{\eta_{tot}} \sum_{s \epsilon S} (\tau_l^s \cap p_l^s) \tag{3.2}$$

where τ_l^s are true labels for each image of a glass slide $s \in S$, p_l^s are the predicted labels for each image of a glass slide $s \in S$, η_s is the number of images for selected primary site $s \in S$, and η_{tot} is the total number of test images in S.

3.6.2 F1-Score

The details provided in the following section are adapted from [84, 85, 86]. F1-score is the weighted mean of precision and recall and is used to evaluate model performance based on uneven class distribution. Precision can be defined as "the number of true positives divided by the sum of the numbers of true positives and false positives", and recall can be defined as "the number of true positives divided by the sum of the numbers of true positives and false negatives" [84]. Considering the classes of datasets used in this thesis, for all datasets, the class distribution in the test split is not balanced. Hence, F1-score derived in Equation 3.3 can provide a better (aggregated) insight into the precision and recall of the trained classifier. The precision value is based on true predictions and actual labels, while recall assesses the true predictions based on all predicted labels. Hence, F1-score is a composite metric used to analyze model characteristics based on precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}. \tag{3.3}$$

3.7 Summary

This chapter provided detailed descriptions of three data repositories, the data distribution among each repository, and the metrics used for model evaluation. For better analysis of the effect of image modality, KIMIA MAG-5 was used. The BreakHis dataset was tested for

low-resolution capability of the developed models, and the final evaluation was conducted on the novel OMAX dataset composed of microscopic snapshots with various primary sites. Inference on the effects of image modality, primary site, and image resolution was obtained from this study. Further, Chapter 4 discusses the techniques and implementation of knowledge extraction methods from the discussed dataset and their application in building a classifier to recognize the magnification levels of test images.

Chapter 4

Characterization and Classification of Microscopic Snapshots

"Science can amuse and fascinate us all, but it is engineering that changes the world."

– Isaac Asimov

4.1 Introduction

THIS chapter discusses the methodologies implemented for designing an end-to-end model capable of discriminating the features of handcrafted algorithms and CNN models with the ultimate goal of recognizing magnification levels. An outline of the complete workflow is discussed in Section 4.2. Image analysis algorithms have a range of hyperparameters. Changing these parameters directly affects the performances of algorithms. Therefore, changing hyperparameters significantly affects the feature extraction and classification performance of all models. This chapter presents the results of experiments conducted using all three datasets: KIMIA-MAG-5, BreakHis, and OMAX. For the KIMIA-MAG-5 and OMAX datasets, the traditional approach of a train–test split was used, whereas for BreakHis images, the folds provided in the literature were implemented to compare the published results [83]. The results for each dataset were analyzed in a three-phase manner: i) analysis of classification algorithms across all magnifications, ii) individual magnification

analysis, and iii) significance of individual organ/patient slides at individual magnification levels. Answers to all the questions posed in the previous chapters are presented in this chapter, along with discussions and associated findings.

4.2 Proposed Approach

In this section, the magnification prediction workflow and its sub-blocks for processing images at various magnification levels are explained. Figure 4.1 shows a complete block diagram of important blocks necessary for processing images and classifying the magnification level.

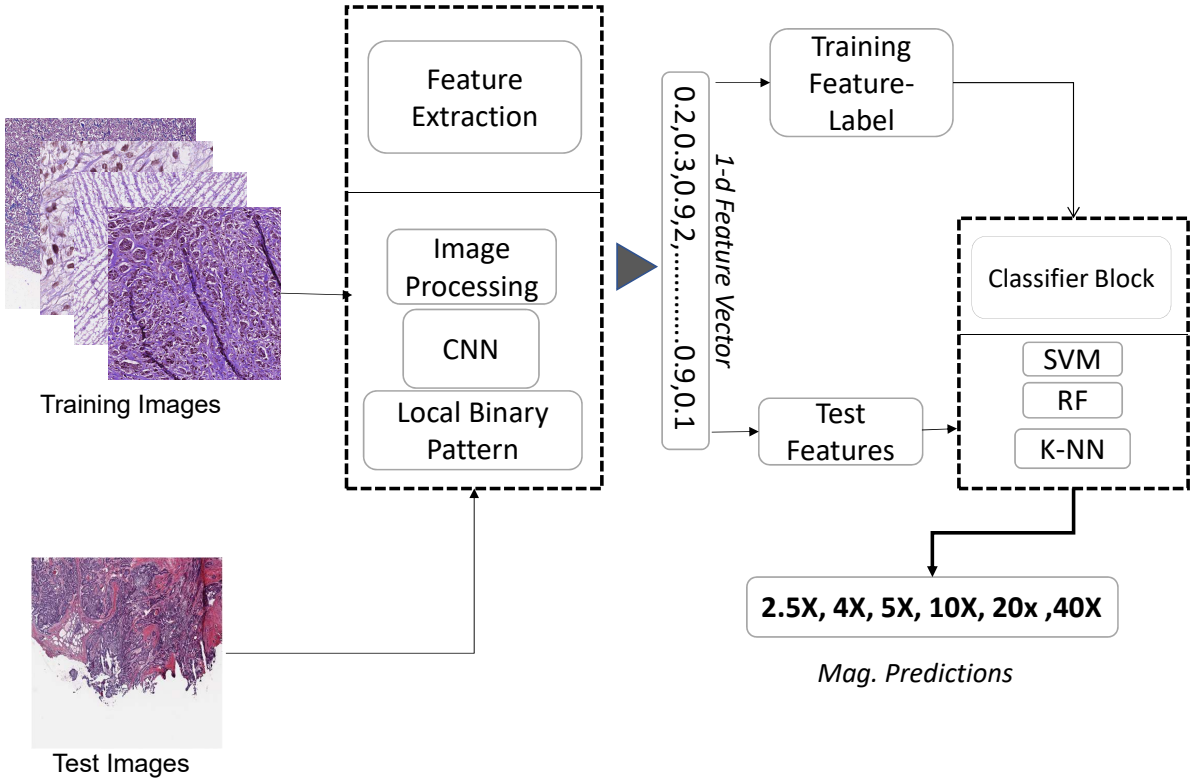


Figure 4.1: Overview of the magnification experiments (from left to right): input of train–test images followed by feature extraction block and image classification.

The diagram shows three primary processing blocks: i) pre-processing block, ii) feature extraction block, and iii) classification block. Pre-processing techniques are specific to the

type of algorithm used (i.e., handcrafted or CNN-based). For handcrafted features, pre-processing constitutes the conversion from the **RBG** color scheme to **gray-scale** images. For deep embeddings, images are normalized by dividing all the pixel values by 255 (2^8-1), which limits the range of intensity values to $[0, 1]$ instead of the original $[0, 2^8 - 1]$. These pre-processed images are provided as input to the feature extraction block. Depending on the type of model used, a 1-dimensional feature vector representing the input image is obtained by either applying a forward pass on CNN models or applying the mathematical function $f(x)$ of the handcrafted feature extraction algorithm. The features generated from the feature extraction block are attached with their associated labels to form a feature–label pair and then passed to the classification block (all the classifiers used in this thesis are described in Section 2.4 to train a classifier and conduct prediction of the test images. The classification block is tuned by adjusting the model-specific parameter values to achieve the best possible classification results.

4.3 Requirements

Access to the three datasets was required to conduct the experiments. All aspects of the datasets can be found in Chapter 3. For conducting the experiments, access to adequate computational resources enabled by the Python environment was required.

4.3.1 Software and Libraries

The code for all experiments was written in Python 3.7, and the virtual environment for conducting all experiments was set up using Anaconda Distribution[1]. The main advantage of using this distribution setup is the default installation of necessary machine learning and data science libraries. A list of essential libraries used here is as follows:

1. Numpy [87]: Popular scientific computing library used for numerical calculations in Python. Matrix calculations are integral to classification algorithms like SVM, RF, and k-nearest neighbors.

2. Pandas [88]: The Pandas library provides fast, flexible, and expressive data structures designed to work with "labeled" data. All image directories and their labels are stored as a Pandas dataframe.

[1] https://anaconda.org/

3. Sci-kit Learn [84]: A prominent library for implementing machine learning algorithms in Python. In this thesis, SVM, RF, and K-NN were implemented using this library based on the train–test split criterion and confusion matrix calculation.

4. TensorFlow [89]: This library was specifically designed for implementation of neural networks. In this thesis, it was used extensively for feature extraction and training neural network classifiers.

5. Keras [90]: Keras is an open-source neural network library capable of running on top of TensorFlow. It enables fast experimentation with deep neural networks and focuses on being user-friendly, modular, and extensible. Feature extraction using CNN models used Keras library in this study.

4.3.2 Resource Configurations

The experiments conducted in this thesis are computationally expensive because they process a large number of histopathology images. Furthermore, the implementation of CNN models for feature extraction and fine-tuning requires GPU use for faster processing. Experiments requiring dynamic memory of more than 31 GB were conducted on a server hosted by KIMIA Lab, University of Waterloo, while other experiments were conducted on a PC provided by KIMIA Lab, University of Waterloo. The CPU-intensive tasks were related to preparing images for training, test, and feature extraction algorithms, such as LBP and VGG, were conducted on the Personal Computer (PC), while feature extraction of DenseNet, ResNet, and EfficientNet and fine-tuning were conducted on the GPU server. The PC has an Intel Core i9 7900x @3.30 GHz CPU, 31.7 GB RAM, and an Nvidia GeForce GTX 1080 GPU. The Kimia Server had 298 GB of shareable RAM available with 4 GPUs (Nvidia Tesla V100).

4.4 Experiment Series 1 (KIMIA-MAG-5)

This section presents the model evaluation conducted using images from the repository *KIMIA-MAG-5* described in Section 3.3. In this series, performance comparison of the handcrafted feature extraction algorithm LBP and deep embedding features using pre-trained CNN models was conducted. For LBP, three combinations of the parameters radius r and neighbors p and three pre-trained CNN models were implemented for the feature extraction process. Features from all algorithms were classified using three classification algorithms, as discussed in Section 2.4, and they were implemented using the Sci-kit learn

library [84]. The images were split based on the primary site, with 80% for classifier training and the remaining 20% as test images for generating predictions from the trained classifier. For these experiments, all images were used in their original dimensions (i.e., 800×800); for LBP, no pre-processing was applied, while for the CNN, all images were normalized before passing to the model.

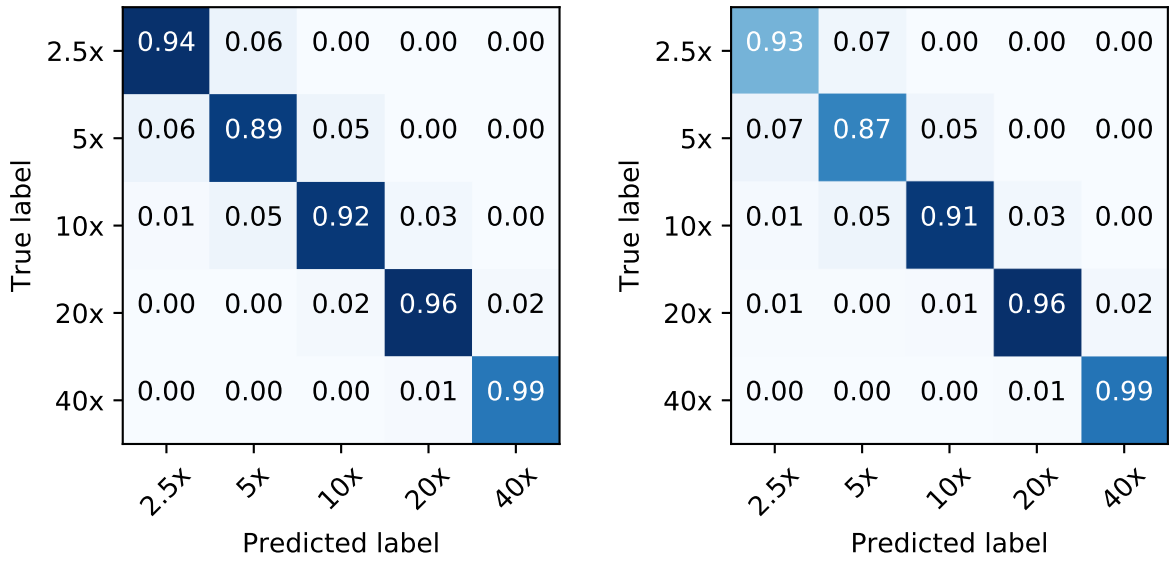
The aim of this experiment series was to answer the primary question *"Is magnification classification of histopathology images possible?"*. In total, eighteen experiments were conducted in this series, consolidating all feature extraction and classification frameworks explained in Section 4.2.

Magnification Prediction – The results associated with the magnification prediction of test images from the KIMIA-MAG-5 dataset are reported in Table 4.1, showing the total accuracy and F1-score for all 18 experiments. In terms of analyzing the domain of feature extraction, a large discrepancy in performance was seen between handcrafted features and deep embeddings. A contrasting difference of 30% in total accuracy was evident between the best-performing handcrafted feature LBPγ_3 and deep descriptor ResNet. It is clear from the results in Table 4.1 that the LBP algorithm was unable to discriminate between magnification variations, whereas deep embeddings were. For handcrafted features, the best-performing model had a total accuracy of 63.28% and F1-score of 0.62 from LBP features with parameters $r = 3$ and $n = 24$ when classified using RF. In terms of deep descriptors, all CNN models were capable of extracting semantic information from magnification-varied images. The top performing pipeline of feature extraction and classification algorithm achieved an accuracy of 93.26% with an F1-score of 0.94 by implementing ResNet151 features with SVM.

Features	SVM		RF		KNN	
	Acc %	F1	Acc %	F1	Acc %	F1
$LBP_{\gamma 1}$	58.36	0.54	**62.90**	**0.62**	54.32	0.52
$LBP_{\gamma 2}$	57.89	0.56	**60.64**	**0.59**	55.72	0.53
$LBP_{\gamma 3}$	62.32	0.60	**63.28**	**0.62**	59.95	0.57
ResNet	**93.26**	**0.94**	82.59	0.84	87.06	0.88
DenseNet	**91.96**	**0.93**	87.13	0.88	88.08	0.89
EfficientNet	**90.96**	**0.92**	86.39	0.88	86.02	0.87

Table 4.1: Total accuracy (T_{acc}) in terms of percentage (%) and F1 score for test split images for three variations of the LBP algorithm:γ_1 for $r = 1$ and $p = 8$, γ_2 for $r = 2$ and p= 16, γ_3 for $r = 3$ and p= 24, and three pre-trained neural networks applied to the KIMIA-MAG-5 repository

Magnification Level Analysis – To understand the misclassifications that occurred at individual magnification levels, an analysis of individual magnification levels was attempted. Figure 4.2 shows confusion matrices from the top two experiments in terms of model performance. An independent analyses of the results of the SVM classification using i) ResNet and ii) DenseNet features are shown in Figure 4.2a and Figure 4.2b, respectively. It is evident from both confusion matrices that the 40x magnification level is the best category for identification, as it was classified correctly 99% of the time. All false predictions for the 40x were categorized in its adjacent class of 20x, and this situation is analogous for all magnification levels. The least generalized results were achieved while testing 5x magnified images; 6% of images were labeled as 2.5x and 5% at 10x. Moderate results were achieved for images from classes of 2.5x, 10x, and 20x.



(a) ResNet features with SVM

(b) DenseNet features with SVM

Figure 4.2: Confusion matrices for analysis of % error in individual classes of magnification levels present in KIMIA-MAG-5. a) Results for true labels on y-axis compared to predicted labels on x-axis for ResNet features. b): Results following same chronology with DenseNet features.

Primary Site Analysis – This section aims to elucidate the influence of primary site (i.e., primary organs from which tissues are taken) on the classification of magnification levels. Figure 4.3 shows the misclassification percentages at each primary site calculated using the total misclassification results aggregated from 5-fold cross-validation. Analyzing

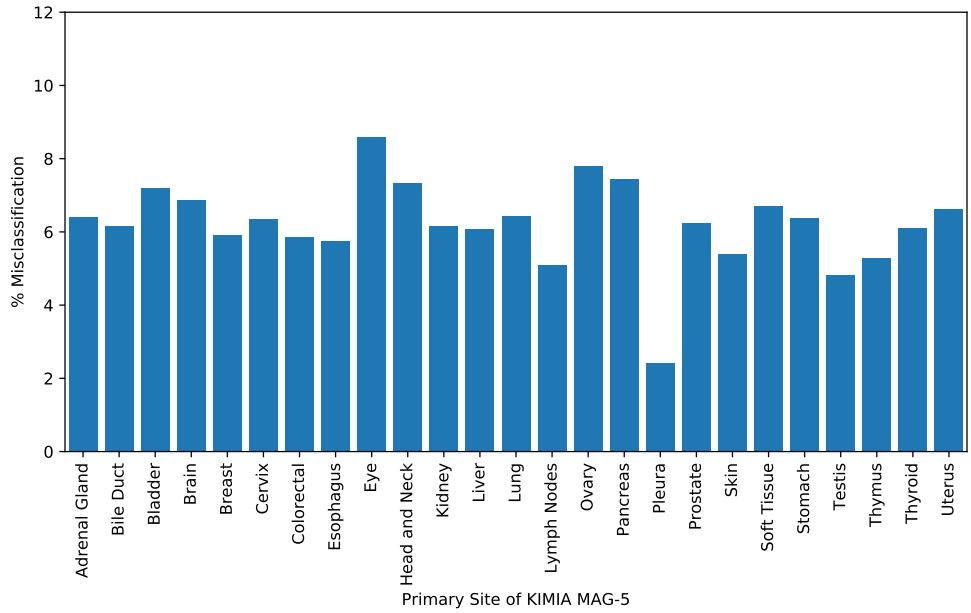


Figure 4.3: Misclassifications for each primary site.

the final results, it is evident that images from pleura were easiest to classify. Eye, head and neck, ovary, and pancreas were the primary sites with the highest misclassification rates. For pleura, only 5 false predictions occurred from a total of 207 images; contrarily for eye, 16 false predictions occurred from a total of 163 images, yielding a high ratio of misclassification error in Figure 4.3.

In summary of Series 1, the deep descriptors tend to be more suitable for capturing magnification-specific information from images. The next section presents the experimental findings from the BreakHis dataset [83].

4.5 Experiment Series 2 (BreakHis dataset)

This section provides the experimental results obtained using images from the repository *BreakHis* described in Section 3.4. The previous section showed that handcrafted features are unsuitable for generalizing magnification recognition. Hence, for this series, the performances of deep descriptors were analyzed using the same CNN models used in Section 4.4.

A total of 12 experiments were conducted for this series using only deep descriptors. Along with magnification, the influence of depth of layers in the CNN was evaluated: *"Can*

shallow neural networks compete with deep neural networks?". To answer this question, VGG19 was included in this series of experiments, and its performance was analyzed across all classifiers.

Magnification Prediction – The results for magnification prediction of images from the BreakHis dataset are reported in Table 4.2. The average validation accuracy was calculated using the average across 5-fold cross-validation, while total accuracy, total patient accuracy, and F1-score were calculated using an independent test set. From the results given in Table 4.2, it is clear that the SVM is the best classification algorithm. Comparing feature extraction models, DenseNet features achieved a top test accuracy of 91.50%. For the EfficientNet and VGG models, the best accuracies achieved were 88.75% and 82.12% with F1-scores of 0.89 and 0.81, respectively. With an F1-score of 0.93, we can state that DenseNet features with SVM provide more robust, generalized, and sensitive predictions.

Magnification Level Analysis – The confusion matrix of the end-to-end pipeline composed of DenseNet and SVM is shown in Figure 4.4. Studying the confusion matrix, it seems that the 20x magnification level is the most difficult to classify. For 20x magnification, the classifier model predicted correct class labels with 88% accuracy, 5% misclassification as 10x, and 7% misclassifications as 40x. These findings contradict our common concept that *"highly magnified images are easy to recognize"*. For this experiment series, low magnification levels were easy to recognize. The 4x magnification level achieved a top accuracy of 97% with 3% of images being misclassified as 10x.

Patient Level Analysis – The BreakHis dataset was manageable in terms of test split, enabling analysis of individual misclassifications of each patient's images from the test split. Table 4.3 shows the patient-level accuracy, misclassified images, and total images for each patient slide from the test split. Interpreting the results in Table 4.3, two patient slides were predicted correctly across all magnification levels. For slides SOB_M_LC_14-

Features	SVM			RF			KNN		
	V.Acc	Acc	F1	V.Acc	Acc	F1	V.Acc	Acc	F1
ResNet	91.98	89.84	0.91	84.03	83.00	0.26	81.00	77.40	0.73
DenseNet	**91.47**	**91.50**	**0.93**	86.47	84.61	0.86	80.00	77.97	0.71
EfficientNet	91.14	88.75	0.89	80.00	78.29	0.81	77.12	71.91	0.68
VGG	78.82	82.12	0.81	75.67	78.34	0.79	75.02	76.99	0.69

Table 4.2: Model evaluation using validation accuracy, total accuracy T_{acc}, and F1-score for magnification prediction of test split of BreakHis dataset. The results from the four feature extraction models are listed row-wise across the three columns (classification algorithms).

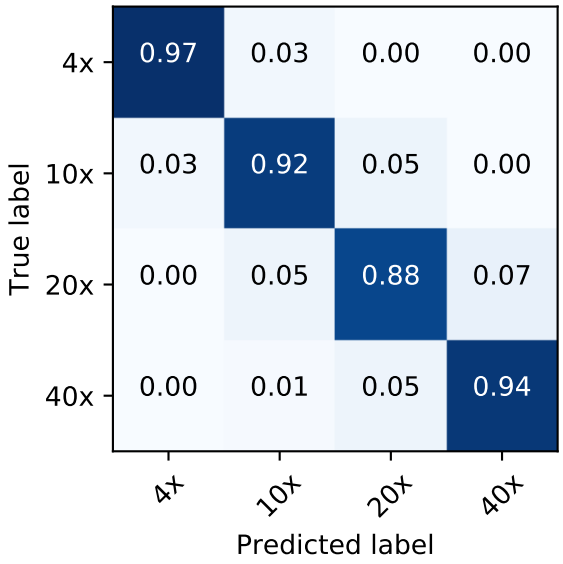
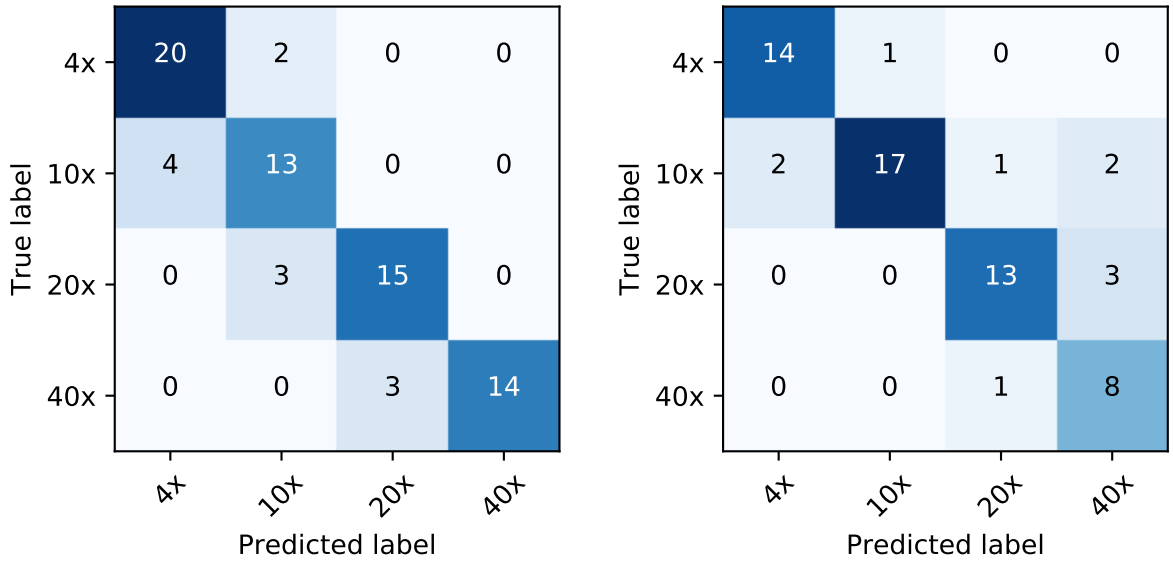


Figure 4.4: Confusion matrix for BreakHis dataset with four magnification levels. Results show that 20x is the most confusing class, whereas 4x is the simplest.

16196 and SOB_M_MC_1418842, a total of 12 and 10 false predictions were respectively obtained across all categories. A complete breakdown of the misclassification is provided in Figure 4.5. The results from Figure 4.5b are surprising because 5 false predictions occurred in the 10x class, which contradicts our previous finding that the 40x class is the 2^{nd} easiest to learn magnification.

Patient Id	Accuracy P_{acc}	μ_f	μ_s
SOB_M_LC_1416196	83.78	12	74
SOB_M_MC_1410147	92.18	5	64
SOB_M_MC_1412773	97.77	2	90
SOB_M_MC_1413413	100.00	0	119
SOB_M_MC_1413418DE	100.00	0	55
SOB_M_MC_1416456	93.25	12	178
SOB_M_MC_1418842	83.87	10	62
SOB_M_MC_1418842D	92.18	5	64

Table 4.3: Patient-level accuracy using primary site accuracy P_{acc} for images from eight patient groups from BreakHis dataset kept as a part of the test split. μ_f is the number of misclassified images, and μ_s is the total count of images for each patient group.



(a) SOB_M_LC_1416196 (b) SOB_M_MC_1418842

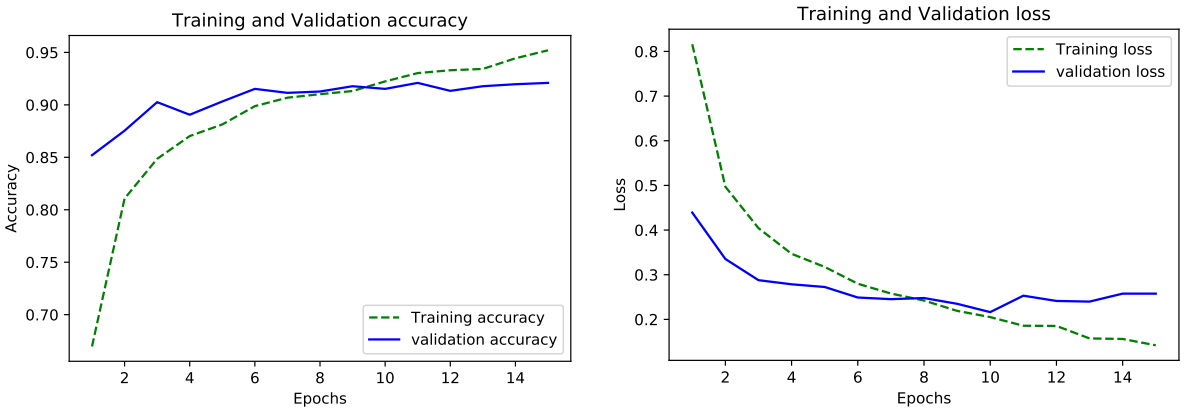
Figure 4.5: Confusion matrix with quantitative analysis of correct predictions compared to false predictions for the two worst-predicted patient slides: a) SOB_M_LC_1416196 and b) SOB_M_MC_1418842.

In summary of Series 2, for each image, individual analysis provides a better understanding of image characteristics. The best pipeline for predicting magnification levels in the BreakHis dataset was found to be DenseNet features classified using SVM. Analyzing the overall model performance provides insight on test data predictions, and investigation on an individual basis yielded surprising results that contradict the findings of Series 1. Hence, individual analysis of testing data is a necessary step in magnification recognition.

4.5.1 Fine-Tuning Deep Networks

Additional experiments on fine-tuning the model architecture for VGG and DenseNet were performed for the BreakHis dataset. The VGG model was cut from the 4^{th} last layer and was attached to fully connected layers for retraining on the BreakHis dataset. A validation accuracy of 91.2% was achieved with a standard deviation of $\pm 1.4\%$; a test accuracy of 90.6% and F1-score of 0.89 were achieved. The modified model was trained using the Adam optimizer with a learning rate of $1e^{-4}$ for 12 epochs with the loss function of categorical cross-entropy. Further training of the model resulted in over-fitting. Figure 4.6 provides

the learning curves of the successful attempt at VGG fine-tuning. In depth analysis of prediction results from test images was not conducted because of its inferiority compared to the pre-trained DenseNet features with the SVM classifier. Furthermore, to compare the results of fine-tuning, the same technique was implemented on DenseNet121. Even after tuning the model parameters, such as the optimizer, learning rate, and momentum, DenseNet121 failed to converge. The test accuracy obtained for fine-tuned DenseNet121 was only 36.7%. A vanishing gradient was observed during the fine-tuning of DenseNet121 from the last block along with fully connected layers.



(a) Training and validation accuracy

(b) Training and validation loss

Figure 4.6: Learning curves based on fine-tuning of VGG19 model: a) Stable flat curve of validation accuracy in blue indicates model is optimally trained. b) Loss function values calculated during each model epoch plotted for training and validation.

4.6 Experiment Series 3 (OMAX)

This section provides the results of experiments conducted using images from the OMAX repository. The main goal of these experiments was to benchmark a dataset and workflow for magnification prediction using microscopic images curated from different organs. The best pipeline from previous experiments on microscopic breast images was achieved using DenseNet features with an SVM classifier. Hence, for this series, three classifiers, SVM, RF, and K-NN, were applied to classify DenseNet features.

The following question is answered by analyzing the results of three experiments conducted in this series: *"Which feature-extraction, classifier model is best for learning mag-*

nification in histopathology domain?" For model testing, 3,328 images were used from 10 organs comprising 15% of the total dataset. The training phase utilized 85% of the dataset with 14,624 images from 54 primary sites.

Magnification Prediction – Model performance on images from the OMAX dataset was analyzed using 5-fold cross-validation and metrics of validation accuracy, test accuracy per organ, and F1-score (Table 4.4). It is evident that the SVM algorithm learned the most optimal differences between magnification levels and generalized the tissue contents from 64 organs. The SVM classifier achieved 87.11% correct predictions, while the variance for model validation was low (1.24) with an F1-score of 0.87, which indicates the high sensitivity and specificity of the trained model. RF and K-NN classifiers did not perform well as they achieved below 75.75% test accuracy and had larger standard deviations, indicating model instability for processing unseen images.

Features	DenseNet		
	V_{acc}	T_{acc}	F1
SVM	**85.77± 1.24**	**87.11**	**0.87**
RF	70.77 ± 5.34	75.75	0.76
KNN	67.14 ± 5.23	72.60	0.72

Table 4.4: Validation accuracy (V_{acc}), total accuracy T_{acc}, and F1-score for deep embeddings obtained using DenseNet model on OMAX test split images and classified using three classification algorithms.

Magnification Level Analysis – The thorough study of different magnification levels indicates the challenges associated with each magnification level. The results from the best workflow model of the SVM classifier presented in Figure 4.7 were studied. By analyzing the confusion matrix, confounding results were seen for 40x magnification. False predictions of 21% of images belonging to 40x were labeled as 4x. This result refutes the statement *"higher magnification levels are easy to classify"*. Analyzing the falsely predicted images in Figure 4.9, it becomes difficult to give a universal statement for all datasets and their magnification level analyses. For the smallest magnification of 4x, misclassification rates of 11% as 10x and 6% as 40x were seen. For this dataset, it seems that the 4x and 40x images show higher misclassification.

Primary site analysis – In total, 10 primary sites were used for analyzing the misclassifications caused by tissue variations. Details regarding the number of misclassified images and accuracy of each primary site are provided in Table 4.5. Recognizing magnification

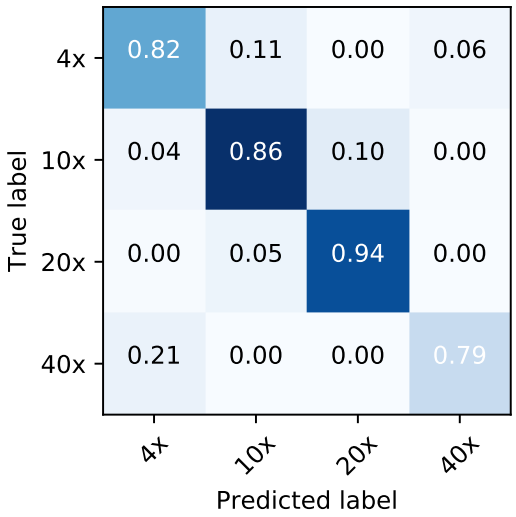


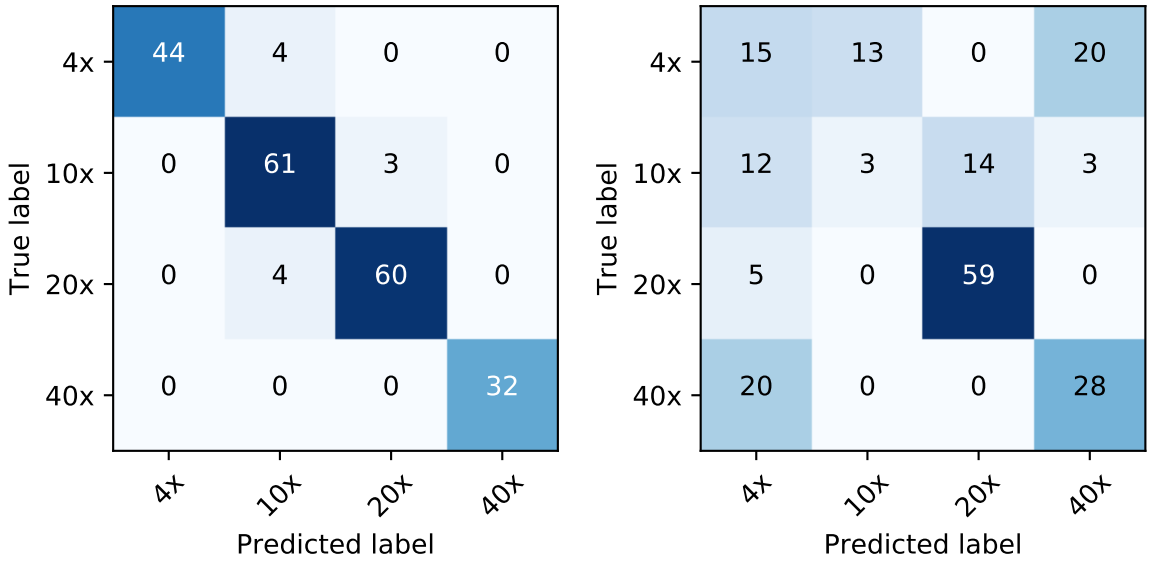
Figure 4.7: Confusion matrix for the OMAX dataset. The model is susceptible to unseen images at 4x and 40x magnifications, leading to high misclassification rates, whereas the mid-level magnifications seem to have been predicted more reasonably.

for images belonging to chromosome and lymph nodes was most difficult for the model, while images from the duodenum, lung, gall bladder, and palatine of the tonsil were easily distinguishable with accuracy scores above 90%. The confusion matrices for the best and worst performances on primary sites are shown in Figure 4.8. The lung images presented in Figure 4.8a had few misclassified images of 4x, 10x, and 20x magnification, while 40x had zero false predictions. In contrast, Figure 4.8b shows the results for chromosome images, where all magnification levels were mostly misclassified.
An analysis of the remaining primary sites is provided in the Appendix B.1.

Misclassification Image Analysis – It was attempted to understand the nature of misclassified images from trained models used for conducting the experiments. Figure 4.9 shows twelve sample test images present in the OMAX data repository. Each image was specifically chosen for analysis to expand the understanding of the complexity involved in this work. Observing these images, it is evident that for lymph node sections, dissimilar staining procedures (trichrome staining) may act as outliers when modeling compatibility directly, resulting in one of the worst performing primary sites in Table 4.5. Further inspection provided information on images from chromosomes that were divergent in terms of homogeneity. These are composed of protein structures and DNA, making the majority area of their images absent of ordinary tissue texture. The remaining images are rich in collagen, nuclei, and cellular matrix. This could be the main reason magnification levels of

Primary Site	P_{acc}	μ_f	μ_s
Adipose Tissue	88.80	43	384
Cardiac$_V alve$	84.65	23	160
Chromosome	54.68	87	192
Doudenum	92.85	32	448
Gall Bladder	90.10	19	192
Lung	**94.71**	**11**	**208**
Lamellar	87.50	32	256
Lymph Node	68.75	80	256
Oesophagus	89.78	36	352
Palatine Tonsil	92.50	66	880

Table 4.5: Accuracy with respect to the primary site P_{acc} on 10 organs from OMAX dataset. Here, μ_f represents the total misclassifications across all magnification levels, and μ_s is the total number of images present for each primary site.



(a) Human lung

(b) Chromosomes

Figure 4.8: Confusion matrices for best and worst performance (Chromosome).

chromosome images are amongst the most misclassified. Analyzing the remaining primary sites, adipose tissue (top left corner) was misclassified as 4x (with the correct label being 40x). Similarly, the palatine tonsil sample (bottom right corner) was misclassified as 4x

(with the correct label being 20x). This could be due to the high amounts collagen and tissue fibers instead of the highly varying texture of dyes observed in the other images. Contradictory results were seen for the human lung. A fine pattern of cartilage or connective tissue with yellow and brown color is seen in the lung images. This color variation might be responsible for confusing the classifier. Similarly, for the remaining images, problems such as irregular textural patterns, loss of focus, and analogous nuclei size for adjacent magnification levels account for the majority of misclassifications.

4.7 Concluding Remarks

This chapter presented the findings from all end-to-end pipelines applied for learning magnification levels from microscopic snapshots. Experimental results showed that handcrafted features are less able to generalize tissue patterns varying at different magnification levels. On the other hand, deep embeddings efficiently represent tissue patterns with different sizes of nuclei, widths of connective tissues, and individual cellular matrix patterns. For the stated problem, the results from Series 1 and 2 demonstrate that low-level and high-level magnified images are generally easy to recognize when the model is trained using H&E staining. In Series 3 with images with less commonly used stains, the models could not generalize. Analyzing various primary sites, it seems chromosomes are most difficult to predict because their images cannot maintain a balance between background and tissue samples due to their extremely small physical size.

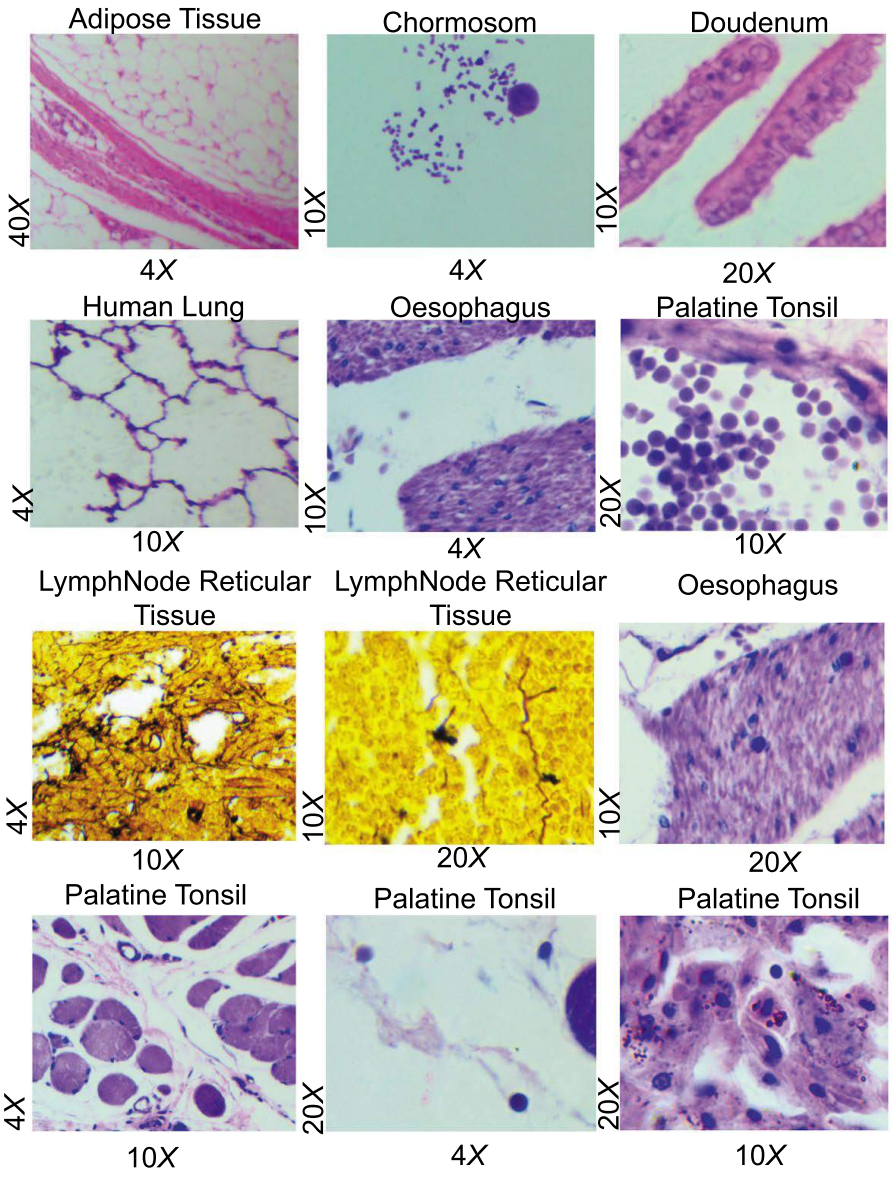


Figure 4.9: Misclassified samples of 12 primary sites from OMAX test samples. Text values on the right side depict actual magnification levels, and the bottom text gives the predicted magnification level by the best-trained model.

Chapter 5

Conclusions and Discussions

"You see, my ambition was not to confound the engineering world but simply to create a beautiful piece of art."

– Kit Williams

The primary goal of this work was to recognize the magnification levels of microscopic images using several feature extraction algorithms from traditional computer vision (i.e., handcrafted features) and deep learning-based pre-trained CNN models. Classification performance was compared using three classification algorithms: K-NN, RF, and SVM. An independent experiment to evaluate the fine-tuning of the neural network was conducted in one series of experiments. With the experiments, we attempted to answer the five questions raised in Section 1.3.

All experiments utilized three datasets: KIMIA-MAG-5, BreakHis, and OMAX. All feature extraction algorithms were able to capture semantic information from the pathology images. Upon classifying these features, different levels of robustness were observed from the classification models. Applying multiple feature extraction algorithms helped in understanding the capability of each algorithm in learning semantically meaningful features responsible for differentiating magnification levels from pathology images.

Experiments on *KIMIA-MAG-5* served as validation of the workflow, along with a comparative study for analyzing the misclassification across modalities. A comparison between handcrafted features, namely LBP, and deep descriptors for three pre-trained CNNs showed that deep embeddings were more capable of learning image characteristics related

to specific patterns, evident at varying magnification levels. In contrast, handcrafted features were poor at acquiring expressive magnification patterns. The best total accuracy of 63.28% and F1-score of 0.62 were achieved for LBPγ_3. For deep features, ResNet features categorized with SVM provided the best results with an accuracy of 93.26% and F1-score of 0.94.

For the *BreakHis dataset*, fine-tuned VGG19 and pre-trained DenseNet combined with SVM achieved very similar results of 90.60% and 91.50% total accuracy and 0.93 and 0.89 F1-score, respectively. The experiment for the fine-tuned DenseNet model from the last convolutional block achieved a total accuracy of 36.7%, which was unacceptable.

The proposed dataset, *OMAX*, is a developing data repository of true microscopic snapshots. The primary motive for developing this dataset was to develop a universal magnification prediction system that can process images from multiple tissue origins. Achieving this goal is challenging because of the tissue complexity seen across various organs. However, in this work, it was shown that state-of-the-art neural networks were indeed able to generalize magnification-dependent features from 64 primary sites. Model evaluation with DenseNet features achieved a total accuracy and F1-score of 87.11% and 0.87 with SVM, 75.75% and 0.76 with RF, and 72.60% and 0.72 K-NN, respectively. In terms of addressing multiple organs, there has been no study that can be used as a benchmark for comparison. Hence, this set of experiments can be considered as a new benchmark for experiments with varying magnifications.

The fine-tuned VGG19 network achieved similar performance to the pre-trained DenseNet121 model for the BreakHis dataset. The trial for fine-tuning the DenseNet model failed as the vanishing gradient phenomenon occurred due to the deep topology of DenseNet121. Therefore, fine-tuning is highly dependent on the available training data and network topology. In terms of primary site analysis, misclassification error in the range of 2–8% was evident in both the *KIMIA-MAG-5* and *OMAX* datasets. As a result, it can be concluded that deep features can robustly represent image patterns across multiple organs and magnifications.

5.1 Future Scope

Existing CAD systems in the field of histopathology have been developed with specific magnification levels (e.g., 20x). This limits the scope of such systems when confronted with images with other magnification levels. If the proposed workflow is used as a pre-processing block for existing systems, the generated results will prove to be more reliable at known magnification levels. In the field of DL, a variety of neural networks are available with accessible computational power and well-annotated datasets. Varying magnification levels

still pose problems for model training; hence, the proposed workflow can be implemented to prepare a magnification-balanced dataset, which will further be used for training a deep neural network from scratch. Currently, no such magnification-robust neural network exists, but it would certainly be a milestone for the DP community.

References

[1] W. H. Organization *et al.*, "Early cancer diagnosis saves lives, cuts treatment costs," *World Health Organization*, 2017.

[2] H. G. Welch and W. C. Black, "Overdiagnosis in Cancer," *JNCI: Journal of the National Cancer Institute*, vol. 102, pp. 605–613, 05 2010.

[3] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, p. 1113, 2013.

[4] M. Babaie, S. Kalra, A. Sriram, C. Mitcheltree, S. Zhu, A. Khatami, S. Rahnamayan, and H. R. Tizhoosh, "Classification and Retrieval of Digital Pathology Scans: A New Dataset,"

[5] P. Carayon and K. E. Wood, "Patient safety," *Information Knowledge Systems Management*, vol. 8, no. 1-4, pp. 23–46, 2009.

[6] J. Guarner and M. E. Brandt, "Histopathologic diagnosis of fungal infections in the 21st century," *Clinical microbiology reviews*, vol. 24, no. 2, pp. 247–280, 2011.

[7] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[8] E. Balogh, B. Miller, J. Ball, and B. on Health Care Services; Institute of Medicine, "The diagnostic process," *Improving Diagnosis in Health Care. National Academies of Sciences, Engineering, and Medicine*, 2015.

[9] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[10] M. Buffington and M. Gates, "Advanced imaging techniques ii: using a compound microscope for photographing point-mount specimens," *American Entomologist*, vol. 54, no. 4, pp. 222–224, 2008.

[11] M. Prochorec-Sobieszek, "Future perspectives of digital pathology," *Nowotwory. Journal of Oncology*, vol. 66, no. 4, pp. 277–284, 2016.

[12] N. Farahani, A. V. Parwani, and L. Pantanowitz, "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives," *Pathol Lab Med Int*, vol. 7, no. 23-33, p. 4321, 2015.

[13] L. Pantanowitz, "Digital images and the future of digital pathology," *Journal of pathology informatics*, vol. 1, 2010.

[14] S. Kalra, C. Choi, S. Shah, L. Pantanowitz, and H. Tizhoosh, "Yottixel–an image search engine for large archives of histopathology whole slide images," *arXiv preprint arXiv:1911.08748*, 2019.

[15] S. Kalra, H. Tizhoosh, S. Shah, C. Choi, S. Damaskinos, A. Safarpoor, S. Shafiei, M. Babaie, P. Diamandis, C. J. Campbell, *et al.*, "Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–15, 2020.

[16] A. Sriram, S. Kalra, and H. R. Tizhoosh, "Projectron–a shallow and interpretable network for classifying medical images," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, IEEE, 2019.

[17] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[18] E. Mercan, S. Mehta, J. Bartlett, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions," *JAMA network open*, vol. 2, no. 8, pp. e198777–e198777, 2019.

[19] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[20] D. Bardell, "The invention of the microscope," *Bios*, vol. 75, no. 2, pp. 78–84, 2004.

[21] X. Chen, B. Zheng, and H. Liu, "Optical and digital microscopic imaging techniques and applications in pathology," *Analytical Cellular Pathology*, vol. 34, no. 1-2, pp. 5–18, 2011.

[22] P. C. Goodwin, "A primer on the fundamental principles of light microscopy: optimizing magnification, resolution, and contrast," *Molecular reproduction and development*, vol. 82, no. 7-8, pp. 502–507, 2015.

[23] A. Kårsnäs, *Image analysis methods and tools for digital histopathology applications relevant to breast cancer diagnosis*. PhD thesis, Acta Universitatis Upsaliensis, 2014.

[24] R. J. Maude, G. C. Koh, and K. Silamut, "Taking photographs with a microscope," *The American journal of tropical medicine and hygiene*, vol. 79, no. 3, pp. 471–472, 2008.

[25] N. J. Desai, B. Gupta, P. Patel, and V. Joshi, "A comparative study of microscopic images captured by a box type digital camera versus a standard microscopic photography camera unit," *Journal of clinical and diagnostic research: JCDR*, vol. 8, no. 10, p. FC23, 2014.

[26] F. Rost and R. Oldfield, *Photography with a Microscope*. Cambridge University Press, 2000.

[27] Z. J. Smith, K. Chu, A. R. Espenson, M. Rahimzadeh, A. Gryshuk, M. Molinaro, D. M. Dwyre, S. Lane, D. Matthews, and S. Wachsmann-Hogiu, "Cell-phone-based platform for biomedical device development and education applications," *PloS one*, vol. 6, no. 3, p. e17150, 2011.

[28] S. Kroemer, J. Frühauf, T. Campbell, C. Massone, G. Schwantzer, H. P. Soyer, and R. Hofmann-Wellenhof, "Mobile teledermatology for skin tumour screening: diagnostic accuracy of clinical and dermoscopic image tele-evaluation using cellular phones," *British Journal of Dermatology*, vol. 164, no. 5, pp. 973–979, 2011.

[29] A. Skandarajah, C. D. Reber, N. A. Switz, and D. A. Fletcher, "Quantitative imaging with a mobile phone microscope," *PloS one*, vol. 9, no. 5, 2014.

[30] L. Pantanowitz, J. Szymas, Y. Yagi, and D. Wilbur, "Whole slide imaging for educational purposes," *Journal of pathology informatics*, vol. 3, 2012.

[31] M. G. Rojo, G. B. García, C. P. Mateos, J. G. García, and M. C. Vicente, "Critical comparison of 31 commercially available digital slide systems in pathology," *International journal of surgical pathology*, vol. 14, no. 4, pp. 285–305, 2006.

[32] M. Thrall, L. Pantanowitz, and W. Khalbuss, "Telecytology: clinical applications, current challenges, and future benefits," *Journal of pathology informatics*, vol. 2, 2011.

[33] D. C. Wilbur, K. Madi, R. B. Colvin, L. M. Duncan, W. C. Faquin, J. A. Ferry, M. P. Frosch, S. L. Houser, R. L. Kradin, G. Y. Lauwers, *et al.*, "Whole-slide imaging digital pathology as a platform for teleconsultation: a pilot study using paired subspecialist correlations," *Archives of pathology & laboratory medicine*, vol. 133, no. 12, pp. 1949–1953, 2009.

[34] N. Stathonikos, M. Veta, A. Huisman, and P. J. van Diest, "Going fully digital: Perspective of a dutch academic pathology lab," *Journal of pathology informatics*, vol. 4, 2013.

[35] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, pp. 331–359, 2013.

[36] N. S. Vyas, M. Markow, C. Prieto-Granada, S. Gaudi, L. Turner, P. Rodriguez-Waitkus, J. L. Messina, and D. M. Jukic, "Comparing whole slide digital images versus traditional glass slides in the detection of common microscopic features seen in dermatitis," *Journal of pathology informatics*, vol. 7, 2016.

[37] B. J. Williams, D. Bottoms, and D. Treanor, "Future-proofing pathology: the case for clinical adoption of digital pathology," *Journal of clinical pathology*, vol. 70, no. 12, pp. 1010–1018, 2017.

[38] J. R. Gilbertson, J. Ho, L. Anthony, D. M. Jukic, Y. Yagi, and A. V. Parwani, "Primary histologic diagnosis using automated whole slide imaging: a validation study," *BMC clinical pathology*, vol. 6, no. 1, p. 4, 2006.

[39] J. Ho, A. V. Parwani, D. M. Jukic, Y. Yagi, L. Anthony, and J. R. Gilbertson, "Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies," *Human pathology*, vol. 37, no. 3, pp. 322–331, 2006.

[40] D. M. Jukić, L. M. Drogowski, J. Martina, and A. V. Parwani, "Clinical examination and validation of primary diagnosis in anatomic pathology using whole slide digital images," *Archives of pathology & laboratory medicine*, vol. 135, no. 3, pp. 372–378, 2011.

[41] T. L. Sellaro, R. Filkins, C. Hoffman, J. L. Fine, J. Ho, A. V. Parwani, L. Pantanowitz, and M. Montalto, "Relationship between magnification and resolution in digital pathology systems," *Journal of pathology informatics*, vol. 4, 2013.

[42] T. Harris, T. Leaven, P. Heidger, C. Kreiter, J. Duncan, and F. Dick, "Comparison of a virtual microscope laboratory to a regular microscope laboratory for teaching histology," *The Anatomical Record*, vol. 265, no. 1, pp. 10–14, 2001.

[43] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Computer methods and programs in biomedicine*, vol. 107, no. 3, pp. 538–556, 2012.

[44] R. J. Zarbo, F. A. Meier, and S. S. Raab, "Error detection in anatomic pathology," *Archives of Pathology and Laboratory Medicine*, vol. 129, no. 10, pp. 1237–1245, 2005.

[45] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," 2016.

[46] K. J. Lim, D. Y. Yoon, E. J. Yun, Y. L. Seo, S. Baek, D. H. Gu, S. J. Yoon, A. Han, Y. J. Ku, and S. S. Kim, "Characteristics and trends of radiology research: a survey of original articles published in ajr and radiology between 2001 and 2010," *Radiology*, vol. 264, no. 3, pp. 796–802, 2012.

[47] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: challenges and opportunities," *Journal of pathology informatics*, vol. 9, 2018.

[48] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," *Computerized medical imaging and graphics*, vol. 35, no. 7-8, pp. 506–514, 2011.

[49] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and structural biotechnology journal*, vol. 16, pp. 34–42, 2018.

[50] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, vol. 7, 2016.

[51] W. S. Campbell, S. M. Lele, W. W. West, A. J. Lazenby, L. M. Smith, and S. H. Hinrichs, "Concordance between whole-slide imaging and light microscopy for routine surgical pathology," *Human pathology*, vol. 43, no. 10, pp. 1739–1744, 2012.

[52] K. L. Weind, C. F. Maier, B. K. Rutt, and M. Moussa, "Invasive carcinomas and fibroadenomas of the breast: comparison of microvessel distributions–implications for imaging modalities.," *Radiology*, vol. 208, no. 2, pp. 477–483, 1998.

[53] M. Veta, P. J. Van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in h&e stained breast cancer histopathology images," *PloS one*, vol. 8, no. 7, p. e70221, 2013.

[54] K. R. Oskal, M. Risdal, E. A. Janssen, E. S. Undersrud, and T. O. Gulsrud, "A u-net based approach to epidermal tissue segmentation in whole slide histopathological images," *SN Applied Sciences*, vol. 1, no. 7, p. 672, 2019.

[55] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, 2017.

[56] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, vol. 35, pp. 489–502, 2017.

[57] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, 2016.

[58] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 284–287, IEEE, 2008.

[59] T. Thomas, "Automated systems comparable to expert pathologists for prostate cancer gleason grading," *Nature Reviews Urology*, vol. 17, no. 3, pp. 131–131, 2020.

[60] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[61] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[62] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218, Springer, 2018.

[63] M. D. Kumar, M. Babaie, S. Zhu, S. Kalra, and H. R. Tizhoosh, "A comparative study of cnn, bovw and lbp for classification of histopathological images," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, IEEE, 2017.

[64] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh, "Comparing lbp, hog and deep features for classification of histopathology images," in *2018 international joint conference on neural networks (IJCNN)*, pp. 1–7, IEEE, 2018.

[65] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International conference on pattern recognition (ICPR)*, pp. 2440–2445, IEEE, 2016.

[66] V. Gupta and A. Bhavsar, "Breast cancer histopathological image classification: is magnification important?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 17–24, 2017.

[67] N. Hegde, J. D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C. J. Cai, M. B. Amin, C. H. Mermel, *et al.*, "Similar image search for histopathology: Smily," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[68] L. Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 4, pp. 249–255, 2003.

[69] H. C. Akakin and M. N. Gurcan, "Content-based microscopic image retrieval system for multi-image queries," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 4, pp. 758–769, 2012.

[70] J. A. Vanegas, J. Arevalo, and F. A. González, "Unsupervised feature learning for content-based histopathology image retrieval," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, 2014.

[71] R. Mormont, P. Geurts, and R. Marée, "Comparison of deep transfer learning strategies for digital pathology," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2262–2271, 2018.

[72] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[73] R. Mehta and K. Egiazarian, "Dominant rotated local binary patterns (drlbp) for texture classification," *Pattern Recognition Letters*, vol. 71, pp. 16–22, 2016.

[74] T. Ahonen, J. Matas, C. He, and M. Pietikäinen, "Rotation invariant image description with local binary pattern histogram fourier features," in *Scandinavian conference on image analysis*, pp. 61–70, Springer, 2009.

[75] M. H. Shakoor and R. Boostani, "Radial mean local binary pattern for noisy texture classification," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 21481–21508, 2018.

[76] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[77] P. Cunningham and S. J. Delany, "k-nearest neighbour classifiers–," *arXiv preprint arXiv:2004.04523*, 2020.

[78] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[79] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[80] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[81] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[82] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[83] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[85] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[86] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology*, vol. 56, no. 1, p. 45, 2008.

[87] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in science & engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[88] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[89] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

[90] F. Chollet, "keras." https://github.com/fchollet/keras, 2015.

[91] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in *Advanced Course on Artificial Intelligence*, pp. 249–257, Springer, 1999.

[92] M. Titford, "A short history of histopathology technique," *Journal of Histotechnology*, vol. 29, no. 2, pp. 99–110, 2006.

[93] R. S. Weinstein, A. R. Graham, L. C. Richter, G. P. Barker, E. A. Krupinski, A. M. Lopez, K. A. Erps, A. K. Bhattacharyya, Y. Yagi, and J. R. Gilbertson, "Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future," *Human pathology*, vol. 40, no. 8, pp. 1057–1069, 2009.

[94] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[95] R. K. McConnell, "Method of and apparatus for pattern recognition," Jan. 28 1986. US Patent 4,567,610.

[96] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[97] M. Zaveri, S. Kalra, M. Babaie, S. Shah, S. Damskinos, H. Kashani, and H. R. Tizhoosh, "Recognizing magnification levels in microscopic snapshots," *arXiv preprint arXiv:2005.03748*, 2020.

[98] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *2009 IEEE 12th international conference on computer vision*, pp. 32–39, IEEE, 2009.

[99] R. D. Huerta, *Giants of Delft: Johannes Vermeer and the natural philosophers: the parallel search for knowledge during the Age of Discovery*. Bucknell University Press, 2003.

[100] A. M. Smith, *From sight to light: The passage from ancient to modern optics*. University of Chicago Press, 2014.

[101] J. A. DeRose and M. Doppler, "Guidelines for understanding magnification in the modern digital microscope era," *Microscopy Today*, vol. 26, no. 4, p. 20–33, 2018.

[102] L. Pantanowitz, J. H. Sinard, W. H. Henricks, L. A. Fatheree, A. B. Carter, L. Contis, B. A. Beckwith, A. J. Evans, A. Lal, and A. V. Parwani, "Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of american pathologists pathology and laboratory quality center," *Archives of Pathology and Laboratory Medicine*, vol. 137, no. 12, pp. 1710–1722, 2013.

[103] S. Mukhopadhyay, M. D. Feldman, E. Abels, R. Ashfaq, S. Beltaifa, N. G. Cacciabeve, H. P. Cathro, L. Cheng, K. Cooper, G. E. Dickey, *et al.*, "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)," *The American journal of surgical pathology*, vol. 42, no. 1, p. 39, 2018.

[104] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[105] S. Kalra, "Content-based image retrieval of gigapixel histopathology scans: A comparative study of convolution neural network, local binary pattern, and bag of visual words," Master's thesis, University of Waterloo, 2018.

[106] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, pp. 582–585, IEEE, 1994.

APPENDICES

Appendix A

Microscope Configurations

Light microscopy works on the principal of light transmission. Magnified objects are viewed using a series of lens. Actual light microscope used for collecting OMAX dataset has an additional attachment to capture snapshots of visual field being observed in real time. Actual system configurations are taken from seller website [1] as follow:

- Total magnification: 40X-100X-250X-400X-1000X-2500X; Eyepieces: widefield - $WF10X/18$ and WF25X; Objectives: achromatic DIN 4X, 10X, 40X (S), 100X (S, Oil); Viewing head: 45 degrees inclined 360 degrees swiveling trinocular; Interpupillary distance: sliding adjustable 2-3/16inch 2-15/16inch (55mm 75mm); Diopter: adjustable on both eyepiece tubes; Nosepiece: revolving quadruple.
- Stage: mechanical stain-resistant double layer, size: 5-1/2inch x 5-1/2inch (140mm x 140mm), translation range: 3inch x 2inch (75mm x 50mm); Photo tube adjustment range: 5/8inch (15mm); Focus: coaxial coarse and fine focus knobs on both sides, rack and pinion adjustment, with tension control; Focusing knob can be locked for observation and transportation; Condenser and diaphragm: Abbe NA1.25 rack and pinion adjustment condenser with iris aperture diaphragm.
- Black palm rest on the base; Illumination: transmitted (lower), replaceable 3W LED light, variable intensity; Metal mechanical components; Power supply: 100V 240V 50/60Hz worldwide range (US and Canada power plug).

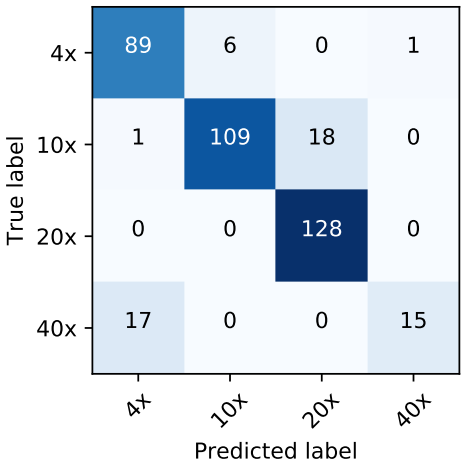
[1] https://www.amazon.com/OMAX-40X-2500X-Trinocular-Microscope-Mechanical/dp/B00SGCEZCS/ref=sr_1_1?dchild=1&keywords=OMAX+40X-2500X+Full+Size+Lab+Digital+Trinocular+Compound+LED+Microscope+with+14MP&qid=1594675372&sr=8-1

- Digital camera: - true color 4096x3288 pixels (14M pixels) - 0.5X reduction lens to get larger field of view - 0.01 mm calibration slide: 1mm/100 division - Frame speed: 1.8fps at 4096x3288, 10fps at 2048x1644, 27fps at 1024x822 - Software compatible with Windows XP/Vista/7/8/10, Mac OS and Linux operating system - Capturing microscope images, recording live video, measuring lengths, angles, areas, editing images - USB cable

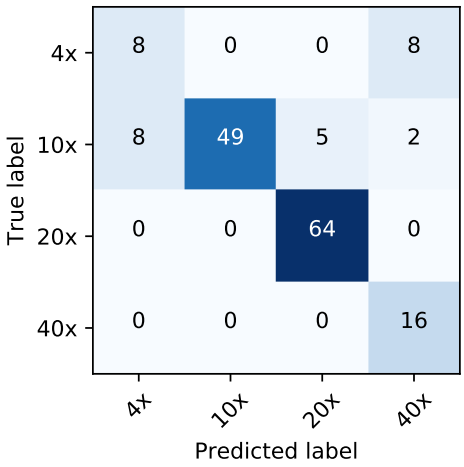
Appendix B

B.1 Additional Results

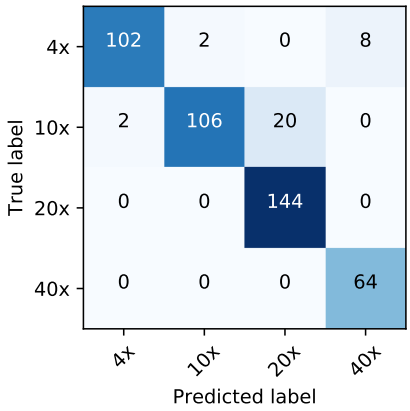
For experiment series 3, total of 10 primary sites are present in test folder. Individual magnification level analysis for best and worst performing primary sites are human lung and chromosome. Results for remaining eight primary sites are provided in this section.



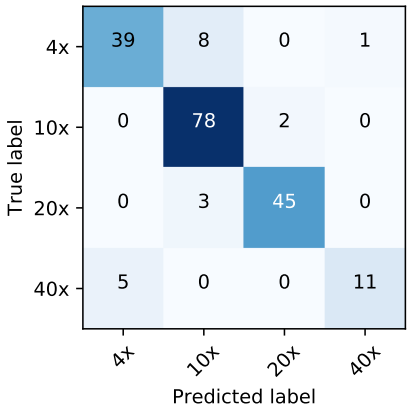
(a) Adipose Tissue



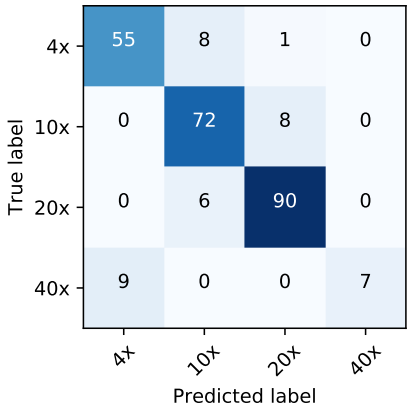
(b) Cardiac Valve



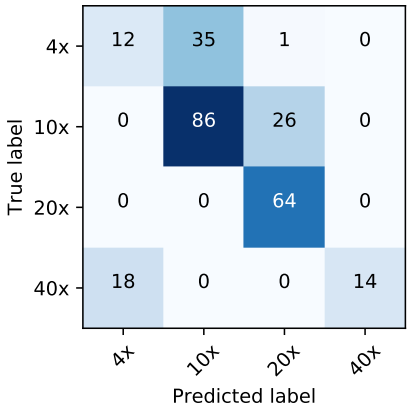
(a) Doudenum



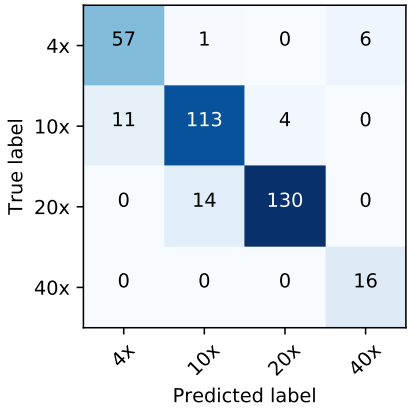
(b) Gall Bladder



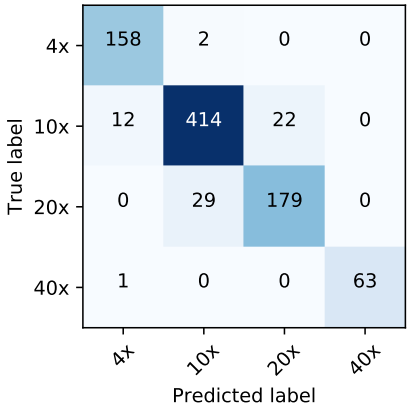
(c) Lamellar Corpuscle



(d) Lymph Node Reticular Tissue



(e) Oesophagus



(f) Palatine Tonsil

Figure B.2: Primary site analysis for 8 primary sites of OMAX dataset