Stain Separation, Cell Classification and Histochemical Score in Digital Histopathology Images

Jingxin Liu

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

June 2018

Abstract

This thesis focuses on developing new automatic techniques addressing three typical problems in digital histopathology image analysis, histochemical stain separation at pixel-level, cell classifications at region level, and histochemical score assessment at image level, with the aim of providing useful tools to help histopathologists in their decision making.

First, we study a pixel-level problem, separating positive chemical stains. To realise the full potential of digital pathology, accurate and robust computer techniques for automatically detecting biomarkers play an important role. Traditional methods transform the colour histopathology images into a gray scale image and apply a single threshold to separate positively stained tissues from the background. In this thesis, we show that the colour distribution of the positive immunohistochemical stains varies with the level of luminance and that a single threshold will be impossible to separate positively stained tissues from other tissues, regardless how the colour pixels are transformed. Based on this observation, two novel luminance adaptive biomarker detection methods are proposed. The first, termed Luminance Adaptive Multi-Thresholding (LAMT) first separate the pixels according to their luminance levels and for each luminance level a separate threshold is found for detecting the positive stains. The second, termed Luminance Adaptive Random Forest (LARF) applies one of the most powerful machine learning models, random forest, as a base classifier to build an ensemble classifier for biomarker detection.

Second, we study a cell-level problem, the cell classification task in pathology images. Two different classification models are proposed. The first model for HEp-2 cell pattern classification comes with a novel object-graph based feature, which decompose the binary image into primitive objects and represent them with a set of morphological feature. Work on cell classification is further

extended using deep learning model termed Deep Autoencoding-Classification Network (DACN). The DACN model consists of an autoencoder and a conventional classification convolutional neural network (CNN) with the two sharing the same encoding pipeline. The DACN model is jointly optimized for the classification error and the image reconstruction error based on a multi-task learning procedure. We will present experiment results to show that the proposed DACN outperforms all known state-of-the-art on two public indirect immunofluorescence stained HEp-2 cell datasets and H&E stained colorectal adenocarcinomas cell dataset.

Third, we study an image-level problem, assessing the histochemical score of a histopathology image. To determine the molecular class of the tumour, pathologists will have to manually mark the nuclei activity biomarkers by assigning a histochemical score (H-Score) to each TMA core with a semi-quantitative assessment method. Manually marking positively stained nuclei is a time consuming, imprecise and subjective process which will lead to inter-observer and intra-observer discrepancies. In this thesis, we present an end-to-end deep learning system which directly predicts the H-Score automatically. Our system imitates the pathologists' decision process and uses one fully convolutional network (FCN) to extract all nuclei region, a second FCN to extract tumour nuclei region, and a multi-column convolutional neural network which takes the outputs of the first two FCNs and the stain intensity description image as input and acts as the decision making mechanism to directly output the H-Score of the input TMA image. To the best of our knowledge, this is the first end-to-end system that takes a TMA image as input and directly outputs a clinical score. We will present experimental results which demonstrate that the H-Scores predicted by our model have very high and statistically significant correlation with experienced pathologists' scores and that the H-Score discrepancy between our algorithm and the pathologists is on par with the inter-subject discrepancy between the pathologists.

List of Publications

1. **Jingxin Liu**, Linlin Shen, Guoping Qiu and Jie Shu. HEP-2 cells classification via novel object graph based feature and random forest. *Biomedical Image and Signal Processing IET International Conference on. (ICBISP)* 2015 [Chapter 4]

2. **Jingxin Liu**, Guoping Qiu, and Linlin Shen. Luminance Adaptive Biomarker Detection in Digital Pathology Images *Medical Image Understanding and Analysis (MIUA) 2016* Procedia Computer Science 90 (2016): 113-118. [Chapter 3]

3. **Jingxin Liu**, Bolei Xu, Linlin Shen, Jon Garibaldi and Guoping Qiu. HEp-2 cell classification based on a Deep Autoencoding-Classification convolutional neural network. *IEEE 14th International Symposium on Biomedical Imaging (ISBI)* 2017. [Chapter 5]

4. **Jingxin Liu**, Bolei Xu, Chi Zheng, Jon Garibaldi, Daniele Soria, Andew Green, Ian O. Ellis, Guoping Qiu. An End-to-End Deep Learning Histochemical Scoring System for Breast Cancer TMA *Submitted to Transaction on Medical Imaging.* [Chapter 6]

5. Zheng Chi, **Jingxin Liu**, and Guoping Qiu. Tuberculosis bacteria detection based on Random Forest using fluorescent images. *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on. IEEE,* 2016.

Acknowledgements

Pursuing a Ph.D is absolutely one of the best decisions I have made in my life. I would like to thank the many people who generously contributed to my doctoral study and research.

First and foremost, I want to express my sincere gratitude to my supervisor, Prof. Guoping Qiu, for the continual support through my whole Ph.D process, for his valuable guidance, patience, and consistent encouragement. He has inspired me to become an independent researcher and helped me realize the power of analytical reasoning. He also demonstrated what a brilliant and hard-working scientist can accomplish.

Similar, I am deeply indebted to my U.K. supervisor, Prof. Jon Garibaldi, who contributed immensely to my research time at Nottingham. I would like to thank Prof. Linlin Shen for valuable advice, feedback, and encouragement. I would also like to thank Prof. Mohammad Ilyas who taught me the histopathological and medical knowledge.

Thanks to my labmates in IDIC vision lab: Bolei Xu, Zhuo Lei, Xianxu Hou, and Ke Sun, for thoughtful discussions and sharing ideas. I also thank my older colleagues: Jie Shu, Chi Zheng, Qian Zhang, Bozhi Liu, and Hao Fu.

I would like to thank the funding support institutes at the University of Nottingham that supported my research, including The International Doctoral Innovation Centre (IDIC) and The Horizon Centre for Doctoral Training (CDT).

I thank with love to my first love, my one true love, my wife, Yiping Jin, for her understanding and love during the past few years. Her unconditional support and encouragement helped me get through this agonizing period in the most positive way.

My parents, Timeng Liu and Li Qiao, receive my deepest gratitude and love for their dedication, unconditional trust, timely encouragement, and endless patience.

Contents

List of Figures

Chapter 1

Introduction

Digital image processing and computer vision techniques have been widely applied to digitized medical images, such as Tomodensitometry (X-ray), Computed Tomography (CT), Magnetic Resonance Image (MRI), Angiography, etc. With the advent of whole slide imaging (WSI) scanners and the increasing demands on histopathologists to interpret the slides, digital histopathology image analysis systems have begun to be developed. This thesis addresses a series of quantitative analyses and assessments, to complement the opinion of the histopathologist.

In this Introduction, we will first review the histopathological background in Section 1.1. In Section 1.2, we will present the motivations behind this thesis. Section 1.3 analyses the challenges that make digital histopathology image analysis difficult. Our contributions to meeting these chanllenges will be presented in Section 1.4.

1.1 Histopathological Background

The word *Histopathology* is a combination of three Greek words: *histo* = tissues, *pathos* = disease, and *logos* = study. It refers to the examination and study of dis-

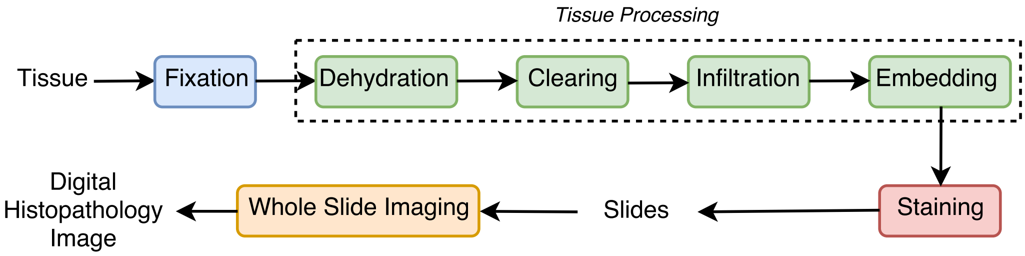


Figure 1.1: The workflow of pathology tissue preparation and digital image production.

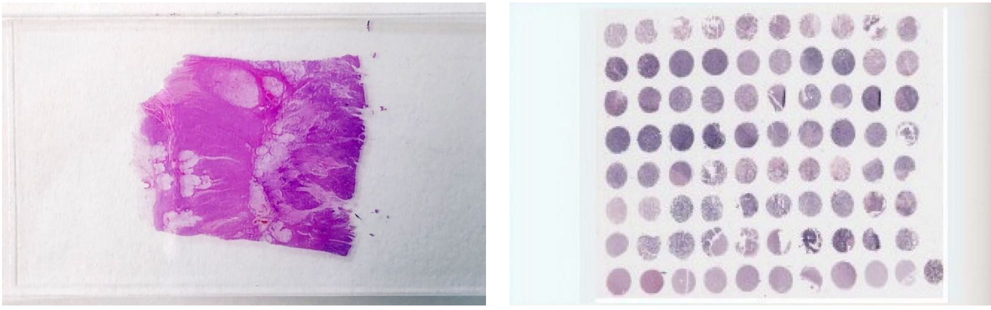
eased cells and tissues under the microscope [5]. A highly trained histopathologist observes the morphology, architecture, and colour information of the cells and tissues in order to make a clinical diagnosis.

A high quality histology slide is the key to this microscopic analysis. A typical pathology workflow of begins with tissue preparation and staining [6], with slide digitization being is added in the modern digital workflow, as shown in Figure 1.1 [7]. Before we proceed to discuss the computer-assisted histopathology image analysis, we will give an overview of digital slide preparation.

1.1.1 Tissue preparation

The tissue specimen is firstly placed in a container with fixatives after being removed from the patient. As the first step of tissue preparation, fixation is to maintain the natural state of the tissue, protecting it from autolysis and putrefaction.

The tissue processing consists of four steps: *dehydration*, *clearing*, *infiltration*, and *embedding* [8]. The aim is to remove water from the tissue and infiltrate it with a medium so that the specimen can be sectioned into thin slices. Alcohol and wax are widely used as the dehydrating agent and embedding medium, respectively. Specifically, in the first step of dehydration, the tissue is immersed in a series of alcohol with increasing concentrations to remove the water. *Clear-*



(a) Whole Slide (b) TMA

Figure 1.2: Examples of pathology glass slides.

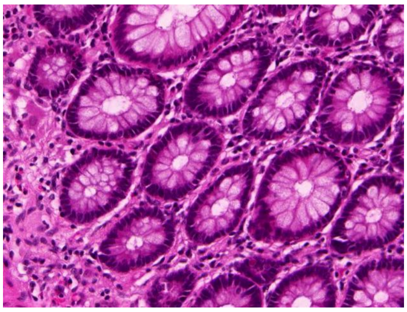
ing is to remove the alcohol with an intermediate solvent (e.g. xylene), since alcohol and paraffin wax are immiscible. Finally, the specimen tissue is placed into a mould with embedding material. The solidified block is then ready for microtomy.

Figure 1.2(a) shows a slide with a section from single paraffin block. In order to achieve a high-throughput analysis, Tissue Micro Array (TMA) has been proposed [9]. In TMA technology, hundreds of small core biopsies are punched from several donor blocks, and transferred to a recipient paraffin block by position in the rows and columns of the array [10] (see Figure 1.2(b)). Such TMA blocks can be used for simultaneous *in situ* analysis of up to 1000 different tissue specimens on the DNA, protein, or RNA level [11].

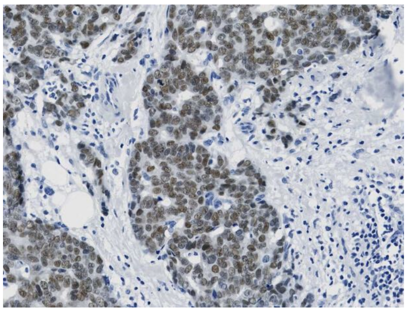
1.1.2 Staining

After fixation and tissue processing, the tissues are dyed with stains to enhance the contrast for viewing under a microscope [12]. In histopathology, there are various methods of staining. In this thesis, three different types of stain methods will be discussed: Hematoxylin and Eosin (H&E), Diaminobenzidene-Hematoxylin (DAB-H), and Immunofluorescence (IF).

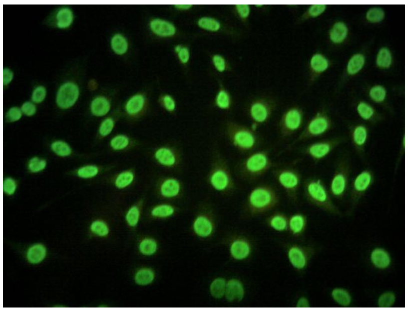
H&E is the routine stain in histopathology laboratories. It has been used for



(a) H&E



(b) DAB-H



(c) IF

Figure 1.3: Images of histopathological stains. (a) Hematoxylin and Eosin, (b) Diaminobenzidene and Hematoxylin, and (c) Immunofluorescence.

over one hundred years and is still essential for recognizing tissue types and morphologic changes [13]. Hematoxylin is a basic dye, with a dark blue or violet colour, and binds to DNA and RNA; eosin is a pink stain that is acidic. In a typical tissue slide stained with H&E, the cell nuclei are stained blue, whereas the cytoplasm and extracellular matrix are pink, as shown in Figure 1.3(a).

Immunohistochemistry (IHC) is a method for detecting targeted antigens (proteins) in tissue sections by labelled antibodies using antibody-antigen interactions [14], and is widely used in cancer diagnosis and biomarker localization [15]. The target antigens are labelled by chromogenic or fluorescence reporters so that they can be visualized through microscopy. Diaminobenzidine is one of the most commonly used chromogenic reporters, which produces a dark brown colouration (positive) against a blue background (negative) counter-stained by Hematoxylin (see Figure 1.3(b)).

Immunofluorescence (IF) makes use of fluorophores as fluorescence reporters to visualize antibodies via a fluorescence microscope [16]. An example of an image is shown in Figure 1.3(c). There are two kinds of IF: direct and indirect. However, these two classes of techniques do not affect the image processing and computer vision methods employed in computer-assisted analysis systems.

Different stains directly affect the methods utilized in CAD systems. The H&E stains all same organelle with same colour, so the CAD systems for H&E stained images would only consider the morphology and texture of tissues; while IHC and IF based CAD systems also need to quantify the stain information such as stain intensity and stain area [17], since the labelled antibodies would stain different colour when antibody-antigen interaction happens.

1.1.3 Digitization

Significant technological developments have led to the advent of whole slide imaging (WSI) scanners. WSI refers to the scanning of an entire glass slide in order to produce a single, high magnification digital image [18].

The resolution of digitalized image is determined by the microscope objective used for the scanning, and is typically at $\times 20$ or $\times 40$ magnification with a spatial resolution of 0.25 micron/pixel or 0.5 micron/pixel, respectively. The captured RGB image can be compressed with either 'lossless' (e.g. TIFF) or 'lossy' (e.g, JPEG or JPEG2000) types [18]. Since the file size is quite large, these images are stored in a pyramid structure with increasing resolution for faster navigation.

1.2 Motivation

As discussed in the previous section, it is a tedious process for histopathologists to prepare a histological slide. Further, the traditional manual examination and analysis have obvious disadvantages in that they are time consuming and can involve subjective mistakes [12].

With the advent of WSI, significant gains in computational power, and dramatic improvements in computer vision algorithms, computer-assisted diagnosis (CAD) systems for histopathology have been highly desired [19]. Specifically, the histopathological CAD systems are designed to give quantitative analyses of histopathology images that can support clinical decision making.

The materials covers different digital tissue slides of most common cancers (e.g. bowel, breast and prostate, etc.) stained by three most widely used stain methods as discussed in Section 1.1.2. The proposed systems range from fundamental aided functions such as region-of-interest (ROI) selection [20, 21] and nuclei detection [22], to high level diagnosis such as automatic biomarker assessment [23]. Compared with a traditional manual analysis, computer-assisted diagnosis promises significant benefits in the delivery of patient care:

- Manual qualitative and semi-quantitative analysis can be transformed into quantitative analysis in an objective way. The number of devastating subjective mistakes would be significantly reduced by different users under the same constraints.
- CAD systems can automatically provide reliable quantitative measurements that can significantly reduce the time taken.
- Digital histopathology images can be stored conveniently, and previous slides can easily be accessed when they are needed. Further, remote access and diagnosis can be achieved.

1.3 Challenges

There are plenty of challenges involved in achieving this desirable goal. Digital histopathology images have many differences from natural images and other medical images. First of all, the size of the whole slide image is extremely large. A 15 mm$\times$15 mm slide with $\times$20 magnification contains 30,000 $\times$ 30,000 pixels. Such a large image size has led to much more complex image pre-processing, and it is difficult to locate the regions of potential interest within such a huge image [24, 25].

Different staining methods bring about various image modalities, and the colour information in histopathology images is significantly important. Colour based segmentation and colour intensity quantification are highly desirable [26, 27, 28]. Nevertheless, there are some challenges faced by colour based CAD approaches. Firstly, the thickness of the tissue section, the variation in stain concentration, and the duration of staining during the histological slide preparation directly affect the stain colour and intensity of the stain. Secondly, the stain colours may have overlapping spectra, which hinders the stain colour separation.

Data-driven machine learning and deep learning techniques have been widely introduced in the medical research community [29]. Nevertheless, histopathology image datasets are always small, due to the complexity and high costs of the data collection [30]. Furthermore, histopathology image labelling requires combining more than one specialist's results, and is extremely time consuming. Thus, it is a challenge for automatic CAD models to get good performance with small datasets. Furthermore, small training set would lead to the problem of over-fitting, resulting in a much better accuracy on the training set than on the test set.

Clinical diagnosis is very highly dependent on the detection, segmentation, and classification of the histological structures, such as the tumour nucleus, the lymphocytes, and glands. However, microscopic images have very complex backgrounds, while the same foreground cellular structures may have distinct variations in shape, size, and texture. It is a big challenge to extract robust features for describing digital histopathology images .

With the increasing application of clinicopathological prognosis, numerous CAD systems have been proposed to support the pathologists' decision making. However, most of existing CAD systems are single task and few researchers consider the image-level synthetic diagnosis. One reason is that image-level applications such as the H-score [31] and HER2 score [32] of immunoreactivity assessments require synthesizing multi-quantitative information, which is a big challenge for automatic CAD systems. In addition, some of the biomarker assessment approaches categorize the IHC stain intensities into different groups [33, 34]. Nevertheless, there is no standard quantitative criterion for grading the stain intensity [35].

1.4 Contributions

In this thesis, we develop different techniques to tackle digital histopathology image analysis problem cover from pixel-level to image-level. This thesis presents the following contributions:

We first consider the pixel-level task of DAB (Diaminobenzidine) stain separation. For immunohistochemical quantification analysis, a careful analysis of the existing DAB stain separation methods has been made. It has been found that the colour distribution of the positive immunohistochemical stains varies with the level of luminance, and previous research with only a single threshold find it impossible to separate positively stained tissues from other tissues, regardless of how the colour pixels are transformed. **Chapter 3** introduces a novel luminance adaptive theory, and proposes a novel DAB detection system using with a luminance adaptive random forest model presenting experimental results to show that the luminance adaptive approach significantly improves the accuracy of the biomarker detection.

This thesis only focus on one cell-level CAD task of cell classification, and two different models are proposed. **Chapter 4** begins by proposing a novel object-graph based feature which decomposes the binarized HEp-2 cell image into primitive objects and represents them with a set of morphological features. A series of experiments has been conducted, and some qualitative results have shown the promise of this approach.

In **Chapter 5**, work on cell classification is further extended using deep learning technique: a Deep Autoencoder-Classification Network. The DACN employs a multi-task learning strategy, which jointly optimizes two loss functions. The experimental results on three different datasets show that the DCAN model outperforms all known state-of-the-art methods, and is particularly effective when the training dataset is small.

Chapter 6 introduces an end-to-end deep learning based framework for histochemical scoring (H-score) assessment of digital TMA images, and both stain colour and tissue morphological information are considered. Instead of dividing the comprehensive task into several small sub-tasks, it creatively constructs two region-attention images that include all necessary quantitative information for estimation of H-score. The final assessment therefore would not be restricted to the performances of sub-tasks. A dual-channel regression network with distribution label augmentation is proposed for histochemical assessment. As far as the knowledge of the present author extends, this is the first work on a TMA-based H-score estimation CAD.

Chapter 2

Literature and Methods

In this chapter, we give an overview of the literature and work related to our proposed digital histopathology image analysis works. We categorize these tasks into three categories: pixel level stain colour separation, cell level application of cell/nuclei classification, and high level comprehensive assessment.

This chapter is structured as follows: it first reviews the related methods for DAB stain separation, and analyses the problems related with them in Section 2.1. Section 2.2 reviews the cell classification literatures, divided into those based on hand-craft features based and those based on deep learning based. Lastly, the literature on high-level assessment of digital histopathology images will be reviewed in Section 2.3.

2.1 Pixel-level Stain Colour Separation and Quantification

Colour is one of the most important features of a histopathology images. In H&E stained tissues, the cell nucleus is stained in blue by the hematoxylin, while eosin (red) the acts as a cytoplasmic stain. Separating the hematoxylin stained pixels can fix the histopathologist's focus on cell nuclei within complex backgrounds [36]. For immunohistochemistry (IHC) with hematoxylin-diaminobenzidine (H-DAB) stains, analysing the distribution and stain intensity of the positive stain (DAB) can be used for biomarker localization or immunoreaction scoring, which are clinically important in cancer diagnosis and prognosis [37, 38]. This thesis will focus on DAB stain separation and quantification because of the greater demand.

To reduce the time needed and the inter-observer variation of manual visual analysis, various computer-assisted approaches have been developed to separate DAB stained tissues in digital histopathology images [39]. These can be divided into three main categories: colour model transformation, colour deconvolution, and mathematical transformation.

Researchers have proposed to convert the original RGB images into other colour spaces to eliminate the correlations in the RGB model [3]. The Hue image from the HSI model has been used by itself [40, 41], while Goto [42] and Kohlberger [43] classified DAB stained pixels by thresholding all three channels. Pham [44] adopted the Yellow channel in the CMYK model, which is believed to have strong correlation with the DAB stain. An OTSU threshold on the YUV colour space has also been used for positive cell extraction [45]. These simple thresholding-based classification schemes are straightforward and easy to implement. However, those colour spaces are not specially designed for DAB stain separation. So the overlap of different stains in the colour spectrum will makes it very difficult to separate the stains completely, thus hindering the accuracy of the stain detection [3].

Colour deconvolution (CD) is perhaps one of the most well known stain separation approaches. It was first presented by Ruifrok and Johnston [1] in 2001. CD was developed based on the properties of light when passing through material according to the Beer-Lambert law. Ruifrok proposed to deconvolve the RGB image into three stain component images. Specifically, for a RGB camera captured digital histopathology image, it is assumed that gray levels in each of the RGB channels are linear with brightness or transmission T, with T being $\frac{I}{I_0}$. By the definition the *transmittance* of material sample is related to its *optical depth* τ and to its *absorbance* A as

$$T = \frac{I}{I_0} = e^{-\tau} = 10^{-A}, \tag{2.1}$$

where I_0 is the intensity of light entering the material sample, and I is the intensity of light detected after passing the material sample. In colour deconvolution related works, $I \in R^{c \times n}$ is the matrix of RGB intensities, where $c = 3$ for RGB channels, and $n =$ number of pixels. Let $M \in R^{c \times r}$ be the stain colour appearance matrix whose columns represent colour basis of each stain such that r is the number of stains ($r = 3$ for colour deconvolution), and $S \in R^{r \times n}$ be the stain density maps, where the rows of which represent the concentration of each stain, then I can be written as follows:

$$I = I_0 e^{-MS} \tag{2.2}$$

I should be first converted to *Optical Density (OD)* values:

$$OD = -log_e(\frac{I}{I_0}) \tag{2.3}$$

Therefore, *optical density* is equal to *optical depth*. OD can be calculated using above equation.

$I_0 = [255, 255, 255]$ for a typical 8 bit RGB camera. Both I and I_0 should be normalized to [0,1] before transform RGB to OD. So the above equation becomes,

$$OD = -log_e(I). \tag{2.4}$$

Consequently, colour deconvolution is given an observation matrix OD, and empirically estimated M, to find stain intensity map matrix S:

$$S = M^{-1}OD. \tag{2.5}$$

Finally, S should be transformed back to RGB space by:

$$S_{RGB} = e^{-S} \tag{2.6}$$

The S_{RGB} is finally image, and each channel represents the corresponding stain. The positive stained tissues are separated by thresholding the DAB component image from S_{RGB}.

A colour deconvolution based stain normalization approach has been proposed in [46], which demonstrates stable performance with scanner variations. Improved methods have also been proposed for unmixing more than three different chromogenic stains [47, 48]. However, DAB does not follow the Beer-Lambert law as the DAB reaction product has a broad and featureless spectrum [49]. Further, precise foreknowledge of the dye colour spectra is required which may be impractical.

Approaches based on mathematical transformation of RGB images are also popular. The *G/B* image was used for DAB selection[50], while Ruifrok [51] presented a *BROWN* image calculated from the RGB image. Two different Blue Normalization (BN) filters [2, 3] have been proposed, and both methods achieve state-of-art performances. All of those methods are developed by calculating the

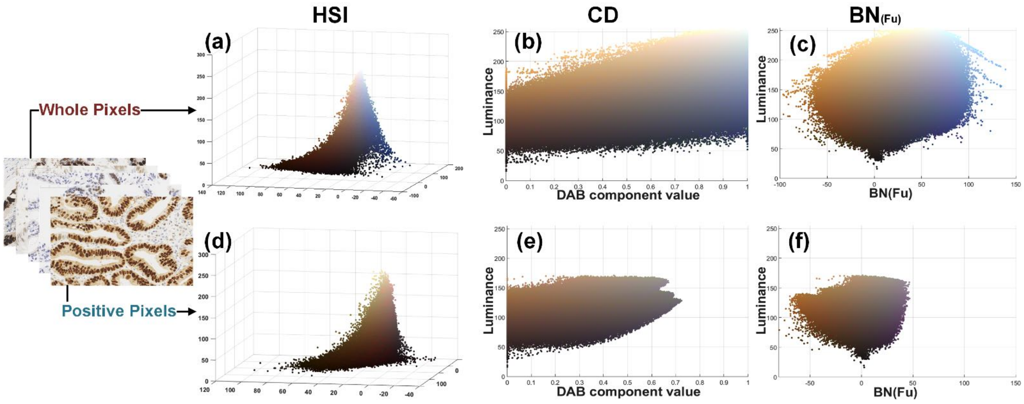


Figure 2.1: Visualizations of pixels of P53 stained histopathology images in the colour spaces of 3 previous methods. Top row: All pixels. Bottom row: Manually labelled positive pixels.

proportion of blue channel as showing below:

$$G/B = \frac{G}{B} \tag{2.7}$$

$$BROWN = B - 0.3 * (R + G) \tag{2.8}$$

$$BN_{Brey} = \frac{255 * B}{R + G + B} \tag{2.9}$$

$$BN_{Fu} = B - \alpha * R - \beta * G, \tag{2.10}$$

where $\alpha, \beta < 1$ are the weights and normally set 0.5 and 0.25 respectively [3]. Similar to colour deconvolution, these methods separate the DAB stain by thresholding transformed single channel images. Since the transformations specially designed for DAB stain separation, the separation performances are better than that of colour space transformations. However, those methods cannot accurately select the positive stained tissues due to the broad colour spectrum of DAB.

Figure 2.1 plots the pixels of hematoxylin-diaminobenzidine (H-DAB) images in the colour spaces of three typical of previous methods mentioned above. Figure 2.1 (a) respectively (d) illustrate the pixels in the HSI cylindrical space for all pixels, respectively, for manually labelled positive DAB stain only. The pixel dis-

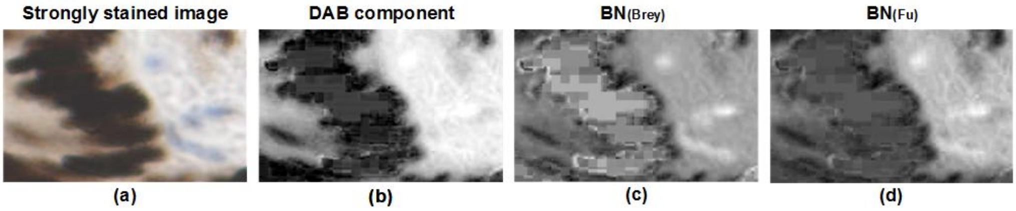


Figure 2.2: Comparison of single channel images of strong stained tissues resulting from three previous methods. (a) The original digital image of strongly stained tissues; (b) the DAB channel from decovoluted image using CD[1]; (c) and (d) transformed single channel images using BN_{Brey}[2] and BN_{Fu}[3] respectively.

tributions taper as the intensity increases. At low intensity, the DAB stain has a significantly broader hue spectrum. Similar plots for the other two colour spaces are shown in Figure 2.1 (b), (c), (e), and (f). Thus, it is obvious that the DAB stained pixel colour has a high correlation with its luminance intensity. However, previous approaches to DAB biomarker detection in digital histopathology images have been based on transforming the original colour channels to form a single channel, to which thresholding is then applied to separate the biomarker from the background.

Figure 2.2 illustrates an example of the resulting single channel images of a heavily stained tissues using the three previous methods mentioned before. The reason is that all of those methods only consider the chroma difference. However, the strongly stained pixels have extreme low values on all RGB channels, which is unable to give significant chroma difference as moderately stained pixels. For instance, the pixel with RGB value of [10,10,10] is extreme darkly DAB stained have same chroma difference with background pixel of [245, 245, 245]. Therefore, it is seen that using simple thresholding cannot completely separate the stains especially if the stained tissues is dark.

Consequently, most previous methods focus on chroma-based separation, which overlooks the influence of the stain luminance. Those approaches have good performance on moderate luminance pixels, which have a significant chroma difference from the background pixels. Nevertheless, the performance is unacceptable on strongly DAB stained pixels with extremely low luminance. Thus, the luminance of pixels, as one of the most important sources of information, should be taken into consideration in work on DAB stain separation.

2.2 Cell-level Application

The cell is the basic unit of live organisms. There are about two hundred different types of cells in the human body, with distinct functional and morphological differences. In pathology, recognizing and distinguishing between different types of cells is the essential work, and histopathologists have sought to exploit the underlying principles of diseases [52]. In this thesis, we only discuss one cell-level task in the histopathology of cell classification. Note that since cell classification is based on recognizing the routinely stained cell nuclei, we will not differentiate the cell and the nucleus in these two applications.

Cell image classification is a typical pattern recognition topic in computer vision, referring to the task of categorizing the cell images into known classes according to the extracted information (features). This thesis will categorize the existing image classification models into two groups: traditional ones based on hand-craft feature, and those based on deep learning models.

2.2.1 Cell classification based on hand-crafted features

In the field of computer vision, traditional image classification models based on hand-craft feature normally contain at least two modules: a feature extraction module and a classification module.

Based on their processing primitives, we categorize them into three groups: pixel-level features, region based local features and image-level features. Early works of cell and nucleus classification utilize pixel-level features. In [53], a set of features, such as the maximum, minimum, and average gray level, as well as the stand deviation of gray, are extracted to classify the cell phase. Nevertheless, the performance of those low-level features is restricted by the quality and the complexity of input images, since the image noise and complex cellular structure directly affect the extracted information.

With the development of computer vision techniques, various region-based handcraft features have been proposed: Histogram of Oriented Gradient (HOG) [54] was proposed for the purpose of shape and object detection, and is widely used for human [55] and vehicle [56] detection; Scale Invariant Feature Transform (SIFT) [57] consists of two parts: keypoint detection and description. Its applications include almost all aspects of computer vision, such as image stitching, object recognition, video tracking, robotic navigation, etc.; Local Binary Pattern (LBP) [58] has been found to be a powerful feature for texture classification.

Beginning in 2012, a series of HEp-2 cell classification contests have greatly promoted the development of cell classification [59]. Numerous hand-crafted models have been introduced. The modified HOG with adaptive robust structure tensors was proposed for HEp-2 cell classification in [60]. Ghosh *et al.* [61] introduced a combination of four different features: SURF, HOG, a texture-based feature, and an ROI-based feature. Nosaka *et al.* [62] proposed Rotation Invariant Co-occurrence Local Binary Patterns (RIC-LBP), which achieved the highest cell level accuracy, 68.54%, in the ICPR2012 HEp-2 cell classification contest. As a improved feature of CoALBP [63] that designed for HEp-2 cells, RIC-LBP is more robust to image rotations retains the descriptive ability. Various LBP features have been introduced for cell classification since then [64, 65], which are proved have a high descriptive ability for HEp-2 cells.

Still more image-level features based on local features have been proposed. One of the most popular methods is the Bag of visual Words (BoW) representation, which creates a codebook using extracted features, and uses a histogram to represent the image. The main difference between the various BoW methods is the codebook generation approach. A BoW feature based on k-means clustering [66] is widely used in whole image classification. To keep the spatial layout information, a discriminative model of spatial pyramid matching (SPM) [67] has been proposed, which firstly segments the image into increasingly fine sub-regions, then computes the histograms of the local features inside each sub-region. There are also other image-level features, such as Fisher vector [68] and covariance matrix representation [69], etc. Those higher level features proved to be the leading strategy in the last decade before the popularization of deep convolutional neural networks [70].

Numerous image-level feature based cell classification approaches have been proposed to improve the classification performance. A typical system combining LBP and a bag of SIFT features, was the winner of the ICIP2013 HEp-2 cell classification contest (74.39%) [71]. Similar work using the Fisher vector [72] has also obtained promising classification results.

Besides the features discussed above, morphological and topological information has also been extracted so as to be used as features for cell pattern classification [73, 74]. Broadly speaking, the difference between the various such systems is their feature selection. LBP and BoW features are the most commonly used features for cell classification. The CAD systems with specially designed features give more accurate classification results. In addition, it is found that fusing different features is able to promote the performance.

Following the feature extraction, a classifier is used to categorize the image into different categories. Although there are various classifiers, support vector machine (SVM) [75] and random forests [76] are the most widely used.

2.2.2 Cell classification based on deep learning

Traditional image classification approaches with hand-crafted features have one common problem: It is difficult to choose or design a set of features which will be well adapted to the whole machine learning system so as to achieve high classification accuracy.

Deep convolutional neural network (CNN) as one of basic architecture of neural networks, has been widely used in computer vision tasks. A typical CNN is built by stacking convolutional and other layers (e.g., pooling, fully connected) hierarchically, and uses the final loss function of softmax with backpropagation to find the optimal features and handle the final classification task.

With the dramatic development of deep learning techniques, various deep neural network based classification models have been proposed [77]. The features learned from the deep learning model have been proven to achieve better classification performance than hand-crafted features [78]. AlexNet with 5 convolutional layers, won the ILSVRC in 2012 [79]. Since then, various convolutional neural network models have been proposed, which have dramatically helped in the natural image classification task [80, 81, 77].

Deep learning methods have also achieved promising results [82, 83] on histopathological image datasets, and very recent cell classification systems have applied the CNN model for feature extraction. A 5-layer CNN model was proposed for classify colon tumour nuclei with H&E stain[84]. Gao *et al.* [85] introduced an 8-layer model with the hyperbolic tangent function for HEp-2 cell classification. Recently, [86] trained CNN model on the mixed HEp-2 image dataset so as to increase its accuracy. Phan [87] applied a pretraining-finetuning approach on the ImageNet and ICPR2012 HEp-2 datasets, and got 77.1% cell level classification accuracy. Li *et al.* [88] and Bayramoglu *et al.* [89] also proposed using a deep learning based model for HEp-2 cell classification.

The published deep learning systems for cell classification have the following characteristics: the CNN architectures are similar and straightforward, with a single cross-entropy loss for the softmax classifier, and the main differences between models are the number and the depth of the convolutional layers. With the development of deep learning techniques, recent proposed CNN architectures for cell classification also become deeper and more complex [90], in order to archive higher accuracy. Nevertheless, the complex CNN model requires larger training set than simple models. Since the size of the training set is fixed for open datasets, researchers have proposed data augmentation, adding other dataset images, or using transfer learning by pre-training the model on other datasets.

2.3 Image-level Comprehensive Diagnosis

Most of the existing literatures on histopathology image analysis carry out single-task works, such as positive stain segmentation as discussed in Section 2.1; cell detection [91] and classification [71]; gland glandular structure segmentation [92] and tubule detection in breast cancer [93], etc. However, the histopathological diagnosis requires a comprehensive assessment of tissue slides. For instance, in breast cancer, metastasis detection and grading in H&E stained lymph node digitized images are an essential part of pN-stage classification [94]; while for confirming tumour histogenesis and subtype (such as the assessments of HER2 expression and histochemical expression), pathologists would take into account both stain intensity and stain percentage in IHC stained histological slides [95].

2.3.1 Step-by-step method

Researchers rarely propose high-level assessment methods for digital histopathology images. Most of the existing image-level comprehensive CAD frameworks are *step-by-step*, which directly follow the assessment criteria by extracting quantitative information from the digital images, which comprise several sub-tasks.

Masmoudi *et al.*[28] proposed an automatic Human Epidermal Growth Factor Receptor 2 (HER2) assessment method, which is a visual examination of cell membrane staining in IHC stained tissue slides to gives a score in categories of {0, 1+, 2+, and 3+} corresponding to no staining, weak, moderate, and strong staining. After extracting the ROI (region of interests), color pixel classifiers were trained to classify the image pixels into three classes: epithelial cell nuclei, epithelial cell membrane, and background. Following pixel classification, epithelial nuclei segmentation and membrane modelling are performed for quantitative feature extraction. Finally, merged features are fed into a minimum cluster distance classifier to classify the slides into HER2 categories.

Trahearn *et al.* [4] established a simultaneous scoring and co-registration process for IHC stained WSI scoring. Similar to Masmoudi's work discussed above, the DAB stained pixels are categorized into four stain intensity groups (unstained, weak, moderate, and strong) by predefined thresholds, after a two-stage registration to co-locate the same ROI for multiple sections. Two different detectors are trained for tumour region and nuclei detection in automated scoring process.

An automatic Ki-67 quantification CAD system is proposed by Shi *et al.* [96] for IHC stained nasopharyngeal carcinoma tissues. After a series of image preprocessing of original digital histopathology image, pixels are clustered into one of DAB and hematoxylin classes based on the local correlation features in multiple colour spaces. Morphological adjusted and watershed based nuclei segmentation are then applied for accurate measurements, followed by positive grading analysis.

All three CAD systems above require a accurate nuclei segmentation and quantification. However, the problems such as the complex cellular structure, stain heterogeneity, overlapping, and blur make automatic nuclei segmentation a challenging problem [22]. The performance still has a certain chasm compared to with pathologists. In summary, those *step-by-step* synthetic CAD systems have a obvious disadvantage that performances are tied to the performance of each sub-task in the whole process.

2.3.2 End-to-end deep learning based method

With the development of deep learning techniques, various deep neural network based CAD models have been published. Deep convolutional networks with deeper architectures can be used to build more complex models which will result in more powerful solutions. Li [97] used an 88-layer residual network for human epithelial type 2 (HEp-2) cell segmentation and classification. *AggNet* with a novel aggregation layer was proposed for mitosis detection in breast cancer histology images [98]. Google Brain presented a multi-scale CNN model to aid breast cancer metastasis detection in lymph nodes[99]. A deep learning based system was proposed for the detection of metastatic cancer from whole slide images. It won the Camelyon Grand Challenge 2016 [100]. A symmetric fully convolutional network was proposed by Ronneberger for microscopy image segmentation [101].

Digital pathology is relative new compared with other types of medical imaging such as X-ray, MRI, and CT. Deep learning, as one of the most powerful machine learning techniques that has emerged in recent years has seen widespread applications in many areas. Yap *et al.* [102] investigated three deep learning models for breast ultrasound lesion detection. Moeskops [103] introduced a single CNN model with triplanar input patches for segmenting three different types of medical images: brain MRI, breast MRI and cardiac CTA. A combination of multi-channel image representation and unsupervised candidate proposals was proposed for automatic lesion detection in breast MRI [104].

Therefore, we believe that a deep convolutional network can be a powerful tool for building high-level medical assessment systems. In this thesis, we proposed a *end-to-end* deep learning-based model for histochemical score assessment.

Only one similar work is found as far as the present author's knowledge, which is a completely data-driven model for tumour proliferation assessment [105]. Several deep neural networks are firstly employed for mitotic feature identification and tumour localization. Deep learning features of WSI (whole slide image) and patch-based tumour are then computed and aggregated with biological features for the comprehensive prognostic assessment.

It can be seen that ROI segmentation is one of the key processes in end-to-end deep learning based comprehensive CAD models. As a typical task in computer vision, numerous deep learning model have been introduced for semantic segmentation. Long *et al.* [106] proposed to transform fully connected layers into convolution layers (FCN) so that traditional classification net could output a heatmap for semantic segmentation. A fully convolutional neural network (SegNet) consisted of an encoder and a corresponding decoder is proposed by Badrinarayanan [107]. U shape net [101] is perhaps the most successful model for biomedical image segmentation. The model has the similar architecture of symmetric encoder and decoder with skip connection strategy, which is widely used in later works [108, 109].

Chapter 3

Luminance Adaptive Biomarker Detection in IHC Stained Images

Accurate and robust computer techniques for automatically detecting biomarkers play an important role in digital histopathology. As we mentioned in chapter 2, the colour distribution of the positive immunohistochemical stains varies with the level of luminance and that a single threshold will be impossible to separate positively stained tissues from other tissues, regardless how the colour pixels are transformed. In this chapter, we propose two novel luminance adaptive biomarker detection methods. The first, termed Luminance Adaptive Multi-Thresholding (LAMT) first separates the pixels according to their luminance levels and for each luminance level a separate threshold is found for detecting the positive stains. The second, termed Luminance Adaptive Random Forest (LARF) applies one of the most powerful machine learning models, random forest, as a base classifier to build an ensemble classifier for biomarker detection.

3.1 Introduction

Diaminobenzidine (DAB) is one of the most commonly used stains in Immunohistochemistry (IHC), which gives a brown colouration (positive) against a blue background (negative) counter-stained by Hematoxylin. Measuring the positively stained areas can provide qualitative assessments of the tissues [38, 110]. Traditionally, this is done manually which is not only labour intensive but also prone to subjective errors. With the increased usage of IHC, it calls for Computer Aided Diagnosis (CAD) systems to support pathologists' decision making. Specifically, we need to develop automatic methods for separating the positively stained tissues from other tissues.

In the literature, various computer-assisted approaches have been developed to separate DAB stained tissues in digital histopathology images [44, 40, 42, 43, 111, 1, 112, 50, 51, 2, 3]. Although significant progress has been made in separation accuracy, there are still several factors affecting the performances of those automatic approaches, such as the concentration of chemicals and the thickness of tissue slides.

In contrast to existing methods that mostly employ simple classification techniques, we firstly propose a Luminance Adaptive Multiple Thresholding (LAMT) method to improve the traditional single thresholding based approaches. Secondly, we develop an advanced machine learning technique termed Luminance Adaptive Random Forest (LARF) classification method for DAB stain segmentation and quantification. Random forest is regarded as one of the most successful machine learning algorithms for developing fast and scalable classifiers [113]. Based on the nature of the problem, our model is an ensemble of random forests, each sub-forest works independently and is adapted to a specific level of luminance. We will present experimental results to show that luminance adaptive approaches can significantly improve the separation performance, and the random forest is ideally suited for biomarker detection.

3.2 Luminance Adaptive Approach for DAB Stain Detection

Based on the analysis of past methods, we propose luminance adaptive approach consisted of multiple thresholding and random forest for DAB stain separation and quantification. We treat luminance as one of the most important information in this work. Pixels are divided into specific luminance intervals before thresholding or classification.

Luminance represents the luminous intensity of images, and human vision has finer spatial sensitivity to luminance. Different from luma, luminance is the weighted sum of RGB components without gamma-correction. In this paper, normalized luminance $l(\mathbf{p})$ of pixel $\mathbf{p} = [R, G, B]$ is calculated according to the ITU-R BT.709 [1]:

$$l(\mathbf{p}) = (0.2126 \times R + 0.7152 \times G + 0.0722 \times B)/255. \tag{3.1}$$

3.2.1 Luminance adaptive multiple thresholding

To improve the performance of single threshold approaches, we use multiple thresholds $T = \{t_1, t_2, ..., t_I\}$ on the final single channel images. Specifically, the transformed pixel $\tau(\mathbf{p})$ using traditional methods is divided into I equal intervals according to the luminance:

$$\tau_i(\mathbf{p}) = \{\tau_i(\mathbf{p}) \in \tau(\mathbf{p}) | \xi_i < l(\mathbf{p}) \leq \zeta_i\}, i = 1, ..., I \tag{3.2}$$

[1] Wikipedia: ITU-R Recommendation BT.709, more commonly known by the abbreviations Rec. 709 or BT.709, standardizes the format of high-definition television, having 16:9 (widescreen) aspect ratio.

where ξ_i, ζ_i are lower and upper boundary of ith luminance interval. Then we threshold the transformed pixel with different values according to its luminance instead of a single threshold. Thus, the threshold t_i is assigned using binary classification as follows:

$$t_i = \underset{c \in \mathscr{Y}}{\operatorname{argmax}} P(c|\tau_i(\mathbf{p})) \tag{3.3}$$

where $\mathscr{Y} = \{c_0, c_1\}$ is the label.

3.2.2 Luminance adaptive random forest

In this work, we treat DAB stain separation DAB biomarker detection as a pattern recognition problem and employ a widely used classification model, random forest, at its core. Considering the stains spectrum spread varies with intensities, we explicitly adaptive the random forests to different luminance levels. We believe this is the first time a powerful pattern recognition technique has been applied to the detection of biomarkers in digital histopathology images.

Feature extraction

For each pixel $\mathbf{p}$, we extract three different types of features to form a 5 dimensional feature: $\mathbf{v}(\mathbf{p}) = (\mathbf{v}_{DAB}, \mathbf{v}_{Cr}, \mathbf{v}_{Cb}, \mathbf{v}_H, \mathbf{v}_S) \in \mathbb{R}^5$. $\mathbf{v}_{DAB}$ is the DAB component of CD image calculated according to pre-set stain vectors [1].We propose to convert the RGB images into YCbCr colour space, and extract Cb and Cr channel information. YCbCr is a method of encoding RGB information, rather than an absolute colour space. Cb and Cr are chroma components of blue-difference and red-difference respectively:

$$Cb(\mathbf{p}) = -0.16874 \times R + 0.33126 \times G + 0.5 \times B; \tag{3.4}$$

$$Cr(\mathbf{p}) = 0.5 \times R - 0.41869 \times G - 0.08131 \times B. \tag{3.5}$$

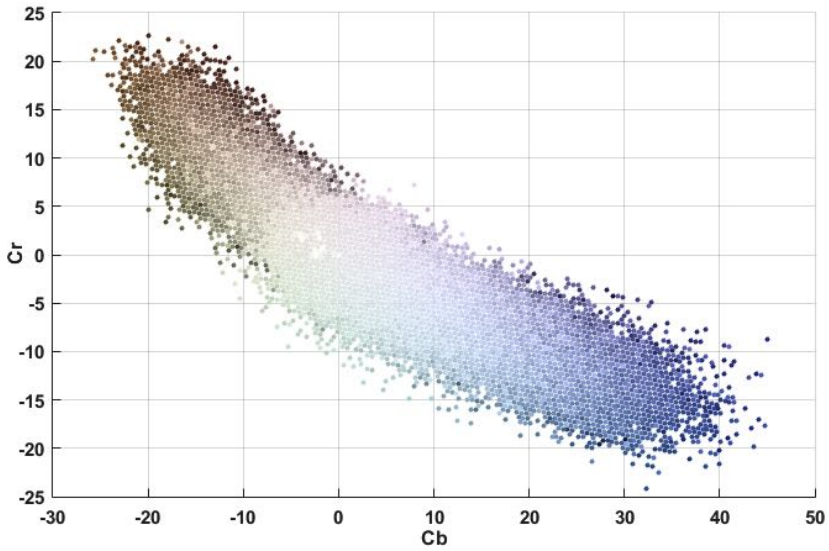


Figure 3.1: The overhead view of H-DAB image pixels in YCbCr space.

The properties of YCbCr are very suitable for DAB separation. The main ingredient of DAB stained brown colouration is *Red*, followed by *Green* and *Blue*; while hematoxylin-stained tissues are *Blue* (see Figure.3.1). Therefore, two different stains can be accurately separated according to Cb and Cr information. In this work, YCbCr values are converted from RGB data. In addition, Hue and Saturation channel values of HSI model are also involved, since the components are correlated better with human perception of color.

Forest training and testing

Similar to LAMT, the training process firstly divides the extracted feature $\mathscr{X}$ into I equal intervals according to the luminance:

$$\mathscr{X}_i = \{\mathbf{v}(\mathbf{p}) \in \mathscr{X} \,|\, \xi_i < l(\mathbf{p}) \leq \zeta_i\}, i = 1, ..., I \tag{3.6}$$

In our method, instead of training single forest, our model is an ensemble of I sub-forests: $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_I\}$. The size of each forest is K, while the whole system has $N = K \times I$ trees. The sub-forest Ψ_i is trained independently with corresponding training set $\mathscr{X}_i$.

The linear split function is used for node splitting, which is setting a threshold on one feature dimension of $\mathbf{v}(\mathbf{p})$. For binary classification, the best situation is that the subset on child nodes are pure containing only positive or negative stained pixels. In this case, we use the widely utilized information gain criterion [114]:

$$Score = \triangle E = -\frac{|\mathscr{S}_l|}{|\mathscr{S}|}E(\mathscr{S}_l) - \frac{|\mathscr{S}_r|}{|\mathscr{S}|}E(\mathscr{S}_r) \tag{3.7}$$

where $\mathscr{S}$ is the training set at split point, while $\mathscr{S}_l$ and $\mathscr{S}_r$ represent the training images contained in the left and right child node respectively. $E(\mathscr{S})$ is the Shannon entropy of $\mathscr{S}$, and $|\mathscr{S}|$ is the number of sample contained in $\mathscr{S}$.

Given a new unlabelled pixel $\mathbf{p}^*$, we firstly extract the luminance $l(\mathbf{p}^*)$ and feature $\mathbf{v}(\mathbf{p}^*)$ information. The luminance information will identify which sub-forest this pixel falls into and the features are fed into the sub-forest for decision making. Specifically, $\mathbf{v}(\mathbf{p}^*)$ is pushed through each tree of forest $\Psi_{\hat{i}}$, if $l(\mathbf{p}^*) \in [\xi_{\hat{i}}, \zeta_{\hat{i}}]$.

3.3 Experiments

3.3.1 Dataset

We conducted DAB separation experiments on the digital slides from two different kinds of DAB-H stained images: Whole Slide Images (WSI) and tissue microarray images (TMA). Both are human colorectal adenocarcinomas slides using the biomarker P53 (nuclear activity), and captured with a $\times 40$ objective lens and scanned using a Hamamatsu scanner (see Figure 3.2). Each type contains 50 images of 1680×1050 pixels. The positively stained nuclei pixels were manually labelled as described in [112].

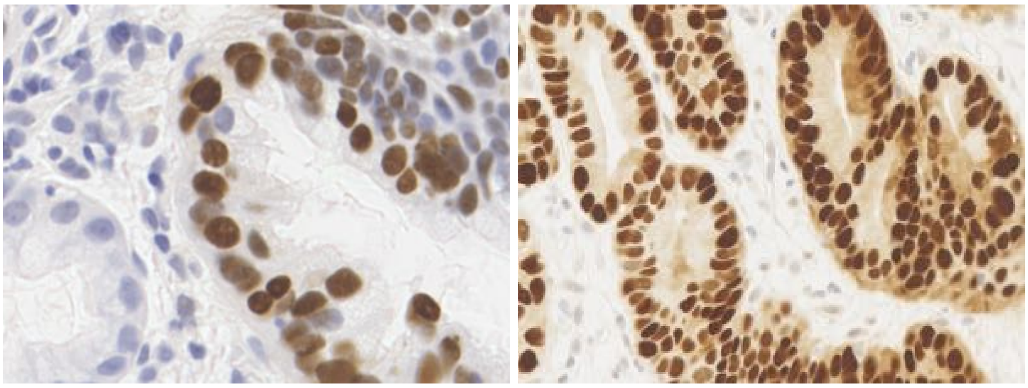


Figure 3.2: Example of DAB-H stained images in dataset.

3.3.2 Evaluation metric

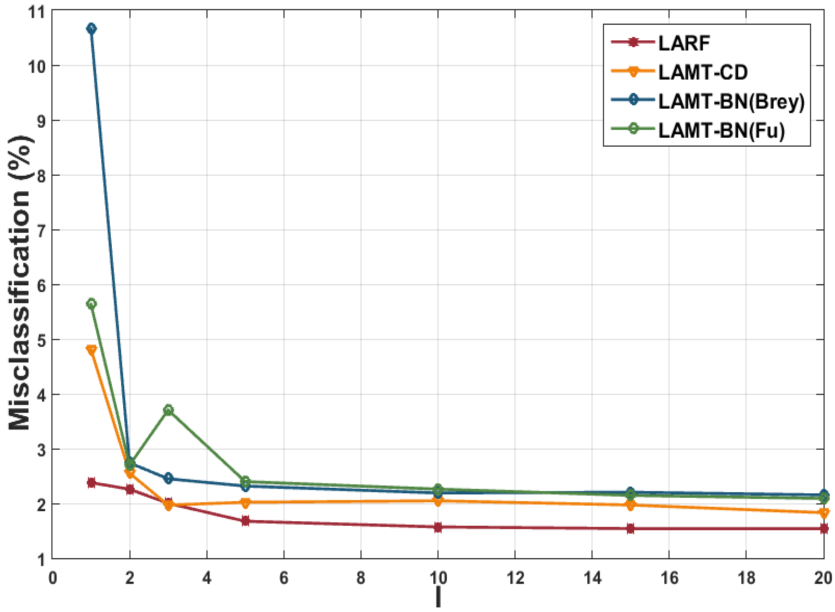
Our results were compared with three state-of-the-art automated separation approaches: CD[1], Brey's BN[2], and Fu's BN[3]. To verify our luminance adaptive approach, we implemented LAMT on those three previous methods. We compared LARF with two conventional random forest models. One of the models used five dimensional features (5-RF) as described in Section 3.2.2, and the other one was trained with six dimensional features (6-RF), which include the luminance as a feature.

A 2-fold cross validation procedure was employed in the experiment, and we repeated the experiment 10 times to test the robustness of the technique. In each 2-fold cross validation, the images were randomly divided into two parts, one part was used for training and the other for testing.

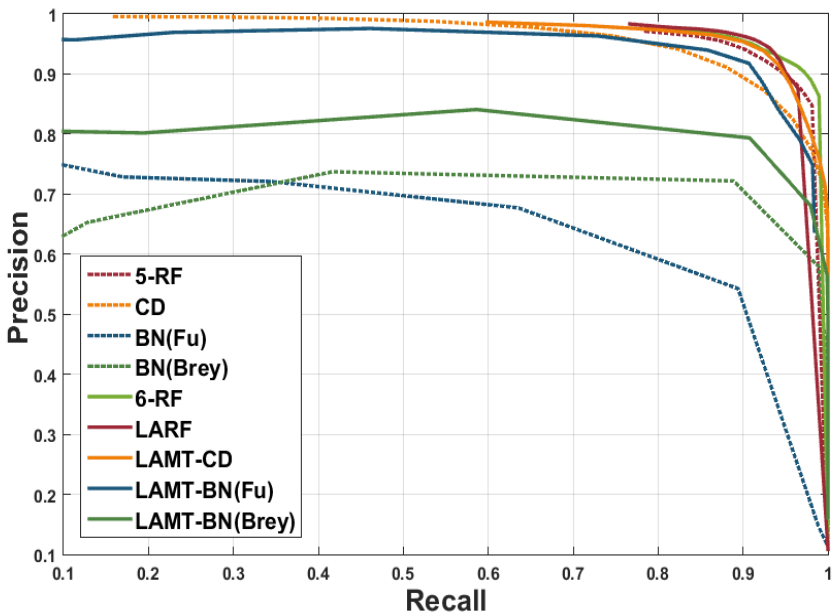
The separating results of the proposed method were compared with manual labelled positive stained nuclei tissues. We employed two criteria to measure the accuracy: misclassification and F1 score.

$$MisClassification = \frac{\sum \text{Misclassified area}}{\text{Total image area}}; \tag{3.8}$$

F1 score is the harmonic mean of precision and recall, where precision and recall are defined as $P = \frac{|TP|}{|TP|+|FP|}$ and $R = \frac{|TP|}{|TP|+|FN|}$ respectively.



(a)



(b)

Figure 3.3: (a) The performance comparison with different numbers of luminance interval I; (b) Precision-recall curve of different methods for DAB-stained tissues separation.

3.3.3 Results

We firstly performed experiments to test the performance of each method for different numbers of luminance intervals I. Figure 3.3(a) shows that the misclassification rates decrease significantly as I increases for all four methods. We found that $I = 10$ worked well and further increase I did not improve the results. It is seen that our method achieves the lowest misclassification rate across all I followed by LAMT-CD. It is seen that Fu's BN gives a higher misclassification rate with $I = 3$ than $I = 2$. One possible reason is that the tolerance of luminance are not suitable for Fu's mathematical transformation method.

Figure 3.3(b) shows the precision-recall curves for each approaches, again, our technique performed the best. Table 3.1 summarises a comparison of misclassification performances (averaged over 10 experiments) of traditional single threshold and LAMT on three previous techniques; while Table 3.2 compares the results of random forest based approaches. It is seen that using the Luminance Adaptive approach, all methods have seen a significant improvement.

Method	Misclassification (%)	F1 Score
CD [1]	4.81	0.812
$\mathrm{LAMT-CD}$	1.83	0.884
$\mathrm{BN_{Brey}}$ [2]	10.66	0.714
$\mathrm{LAMT-BN_{Brey}}$	2.15	0.877
$\mathrm{BN_{Fu}}$ [3]	5.63	0.780
$\mathrm{LAMT-BN_{Fu}}$	2.09	0.831

Table 3.1: Performance comparison of traditional single thresholding method and LAMT method with three previous approaches.

Both 5-RF and 6-RF have 100 tress, and the each sub-forest in LARF also has 100 trees. It is also worth noting that a random forest that included the luminance as a feature (6-RF) also achieved a lower misclassification rate (1.89%) than all previous methods. This demonstrates that random forest can be used as a powerful classifier for accurately detecting biomarkers in digital pathology images.

In this case, the luminance is treated by the random forest as a separate feature and the ways it is used for separating the biomarkers are found through the random forest construction mechanism. Whilst including luminance as a feature worked well, pre-segmenting the luminance and building sub-forests adaptive to the luminance level gave the best performances.

We also compare the computational time of different random forest based models. MATLAB is used as the software environment, and the evaluation computer has an Intel Core i7 3.40 GHz and 16 GB memory. The computational time shown in Table 3.2 is the testing time of a single test image of 1680×1050 pixels. It can be seen that our proposed LARF takes less time than 5-RF and 6-RF. One possible reason is that luminance division reduces the complexity of some sub-forests, which directly accelerates the computation.

Consequently, the result indicates that hitherto ignored luminance is the key for solving the problem of incomplete separation caused by DAB featureless colour spectrum, and packet-based DAB stained pixel selection according to its luminance is simple and efficient.

Method	Misclassification (%)	F1 Score	Computational Time (S)
5-RF	2.38	0.863	5.37
LARF	**1.53**	**0.905**	**4.01**
6-RF	1.89	0.887	5.69

Table 3.2: Misclassification, F1 score, and computational time of random forest based methods.

3.4 Summary

In this Chapter, we have presented a novel quantitative analysis method for H-DAB stained digital images. Our model treats luminance as a very important information for DAB segmentation and trains several sub-forests separately each adapted for a specific level of luminance. We conducted a series of experiments to evaluate the new method. The results demonstrate that misclassification errors can be significantly reduced by including luminance information in the decision making for both traditional methods and for random forest classifier. Luminance adaptive random forest approach is shown to give the best performances.

Our proposed models still have room for improvement. The experiment shows that the misclassification rate would not be improved after luminance interval $I > 10$, and equal division of luminance may not suitable for all DAB separation methods. Therefore, dynamic self-adapting approach will be developed in the future work. In addition, we found that DAB stained pixel labelling is time-consuming, and suffered from subjective mistakes. Unsupervised learning, semi-supervised learning, or online machine learning based models for stain separation task are highly desired.

Chapter 4

Object Graph Based Feature for HEp-2 Cell Classification

In traditional cell classification works, the selection of hand-craft features directly affects the final classification performance. In this chapter, we introduce a novel object-based method to decompose the binary image into primitive objects and represent them with a set of morphological features, after careful observation of Human Epithelial type 2 (HEp-2) cell images. The advantages of this system are as following: robustness against the changes of intensity and rotation; more discriminative information compared to normal morphological descriptors. We evaluate the proposed approach using the publicly available HEp-2 I3A-2012 datasets. The experimental results show that the proposed method achieves comparable performance with the hand-craft feature based methods.

4.1 Introduction

Indirect Immunofluorescence (IIF) detects target anti-nuclear antibodies (ANA) with fluorescent-labelled antibodies so that the target molecule with a cell can be visualized via the fluorescence microscope. The recognition of HEp-2 cells is one of the core challenges for ANA tests.

Unfortunately, the process of IIF is complex and time-consuming. In addition, manual analysis is the primary way to identify the fluorescence pattern, which is heavily dependent on the experience of pathologists [12]. Hence, Computer Aided Diagnosis (CAD) systems are proposed to overcome these limitations, and support the pathologist's decision. Specifically, an automatic classification scheme based on staining pattern is arranged to reduce time cost, and provide a reliable diagnosis based on its quantitative measurements. There has been a growing interest in using CAD for HEp-2 classification since the organization of contests hosted by ICPR 2012 , ICIP 2013, and ICPR 2014. Lots of CAD systems have been introduced since then, which have greatly promoted the development of HEp-2 cell classification.

In spite of the significant progress of HEp-2 cells classification, there is still a large margin for improvement of the performance. In this chapter, we first adopt a series of pre-processing techniques for noise removal and image enhancement. Along with a set of local gradient based textural descriptors, we propose an efficient object-graph based method to extract morphological and shape information, which is specially designed for the I3A-2012 dataset. Finally, the object-graph based feature combined with local texture feature using the bag-of-word model to improve the classification accuracy.

4.2 Methodology

4.2.1 HEp-2 Image

According to the intensity of fluorescence, HEp-2 cell images can be divided into three levels: absence, intermediate, or positive [73]. In this work, we only con-

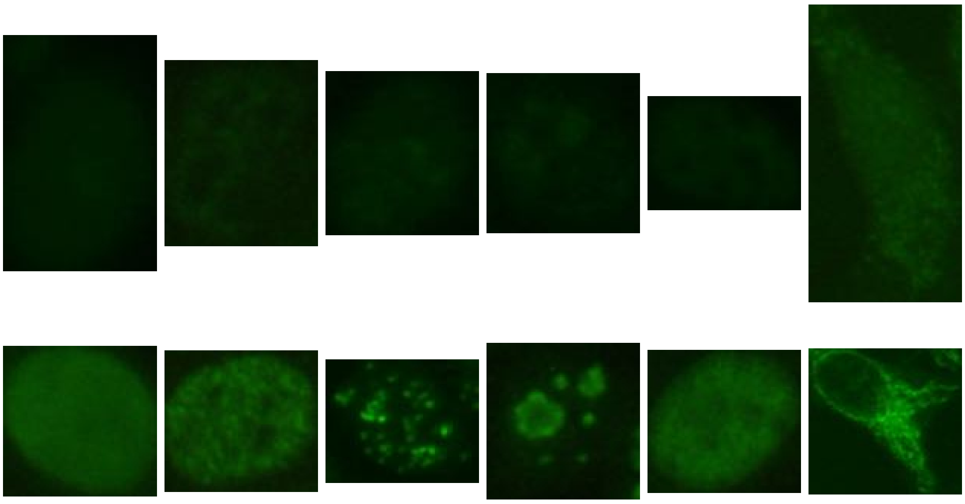
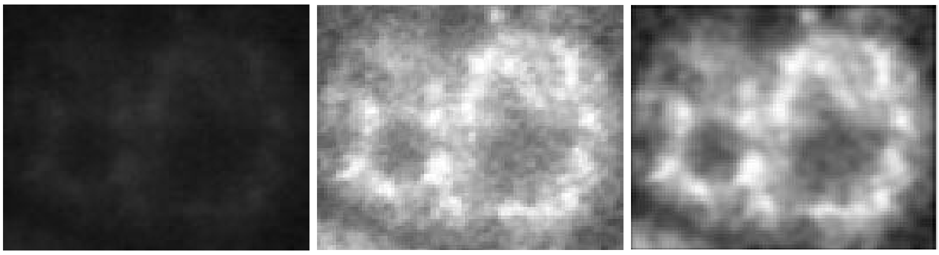


Figure 4.1: Example of HEp-2 cells. The top row illustrate intermediate images, while the bottom row shows positive stained cells. The staining patterns from left to right are *homogeneous*, *coarse speckled*, *centromere*, *nucleolar*, *fine speckled*, and *cytoplasmic* respectively.

sider the I3A-2012 HEp-2 cell images, which is consisted of six most frequent patterns corresponding to different autoimmune diseases[115]:

- **Homogeneous (H)**: a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells
- **Coarse Speckled (CS)**: a coarse granular nuclear staining of the interphase cell nuclei
- **Centromere (C)**: several discrete speckles distributed throughout the interphase nuclei
- **Nucleolar (N)**: large granules in the nucleus, and less than six per cell
- **Fine Speckled (FS)**: a fine granular nuclear staining of the interphase cell nuclei
- **Cytoplasmic (Cy)**: fine fluorescent fibers running over the length of the cell.



(a) (b) (c)

Figure 4.2: (a) is the green channel image of intermediate stained HEp-2 cell; (b) is the image after normalization; (c) is the image after all pre-processing procedure.

Examples of two intensity levels and six patterns are shown in Figure.4.1.

4.2.2 Pre-processing

Intermediate cells have extremely low intensity (as shown in Figure.4.1), which is caused by high dye concentrating in parts of slides. Due to that, the classification task is harder than expected. Thus, we first pre-process all cell images in order to extract reliable information for classification.

Since IIF stained images appear green under a fluorescence microscope, and most information is in the green channel, we only extract the green channel from the original RGB image. Contrast stretching is employed to normalize the image. The processed image brings out improved brightness and many subtle details (Figure 4.2(b)). As texture information is extremely important, the normalized image is firstly filtered with a Gaussian filter for noise removal, then convolved with a Laplacian kernel for image sharpening. Further, the effect on background caused by normalized could be reduced by Laplacian filter. Figure 4.2 compares the original green channel image with the normalized image and the full processed image.

4.2.3 Feature extraction

Object-Graph

The shape and morphological information of HEp-2 have been attracting increasing attention of researchers [73, 116]. In this section, we will introduce the object-based morphological feature extraction method. More specifically, the image would be decomposed into a set of primitive objects and then a set of features according to those objects will be extracted. A similar method is first proposed for gland segmentation [117].

We would first interpret why object graph could have a good performance on HEp-2 cell images. It is obvious that homogeneous, fine speckled, and coarse speckled have circle shapes, while the region with high intensity in centromere and nucleolar can also be replaced by circles. Thus, the stained patterns could be transformed to a set of circle objects.

The object-based method comprises a series of analysis steps. In the first step, we binarize the image using N increasing intensity thresholds. Binary masks and foreground pixels representing the cell pattern could be obtained according to those thresholds as shown in Figure 4.3. The first threshold T_{mid} is calculated by K-means clustering on the histogram of cell images. Two intervals of the histogram are then equally segmented by I background and J foreground thresholds. Thus, the entire threshold set can be defined as:

$$T = [T_{b1}, T_{b2}, \cdots, T_{bI}, T_{mid}, T_{f1}, T_{f2}, \cdots, T_{fJ}], \tag{4.1}$$

while $N = I + J + 1$. Consequently, $N+1$ binary masks can be produced with N intensity thresholds. In addition, we also transform the hole region in binary mask to circles, to distinguish coarse speckled and fine speckled patterns from others. Based on the observation of different binary masks, we the founded that the feature extracted from middle threshold binary images are more discriminative, and middle hole images are also informative for cell pattern recognition. Thus, five binary masks and two hole masks are extracted for circle-transform, i.e.,

$$M = [B_{bI}, B_{mid}, B_{f1}, B_{f2}, B_{f4}, H_{f1}, H_{f2}], \tag{4.2}$$

where B is the binary mask, H is the hole mask, and the subscript denote the threshold index.

Algorithm 1 Circle-Transform Algorithm

Input: I: binary mask; τ: radius threshold
Output: I_{trans}: transformed image

```
function IntCirTrans(I, τ)
    I_loop ← I
    for loop <= 2 do
        [I_trans, R_max] ← CirTrans(I_loop)
        Eliminate small circle using τ
        Eliminate pixels except the circle with R_max in each connected region
        Get I_loop for next loop
    end for
    return I_trans
end function
function CirTrans(I_loop)
    CRN ← number of connected regions
    for i <= CRN do
        pixNum ← pixel Number in i^th connected region
        for j <= pixNum do
            R_max ← 0
            Find the largest circle with radius R_j
            if R_j > R_max then
                Label the pixel in the circle with R_j
            end if
        end for
    end for
    return I_trans, R_max
end function
```

After eliminating non-cell regions using size-based probability, the foreground pixels in binary masks are then transformed into circle objects. Specifically,

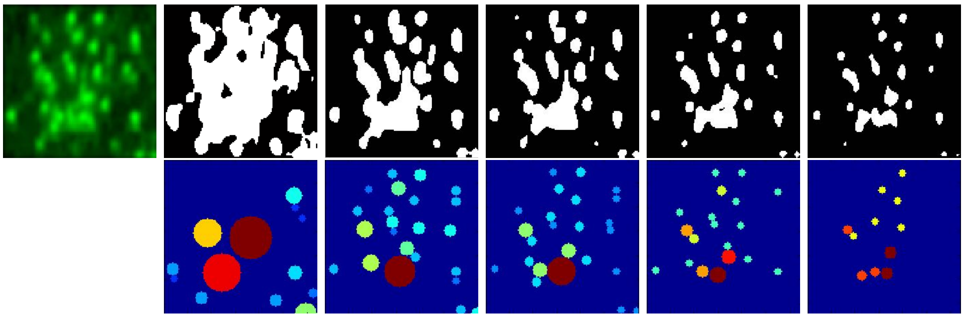


Figure 4.3: Examples of binary images (top) obtained by different thresholds on a centromere HEp-2 cell image, and corresponding circle-transformed images (bottom) extracted using Algorithm 1.

for each positive pixel, we find the radius of the largest circle that contains this particular pixel as the centre of the circle in the independent region. We repeat this process on the remaining positive regions, and eliminate the circle with the radius under an empirically determined threshold τ. The pseudocode of circle transform algorithm is given in Algorithm 1. Figure 4.3 illustrates the binary image and corresponding circle-transformed image.

For each transformed image, we extract morphological properties as below:

- The maximum, the minimum and the average area of the foreground region;
- Average circularity of the foreground connected components;
- The number of circles per connected components;
- Total number of circles;
- The maximum, minimum and average radius of circles;
- the coordinate of the maximum circle.

The circularity is calculated by dividing the area by the squared perimeter: $Cir = \frac{A}{L^2}$. Consequently, a representative vector of size 11 is extracted on each circle-transformed image, which is finally combined as a morphological descriptor of size 77.

Spatial Pyramid Matching

Spatial Pyramid Matching is an extension of the bag-of-words scheme, which represents the image as a sparse vector of occurrence counts of words [118]. Conventional pyramid matching works by computing the weighted histogram intersection using multi-resolution histograms [119]. While SPM is performed in the two-dimensional image space with bag-of-words feature space.

The dense SIFT [57] of 16×16 patches is employed to describe the cell image. For better discriminative power, we propose rootSIFT [120] rather than skipping normalization in conventional SPM. SIFT vectors are first L_1-normalized, and then square root each element, i.e.,

$$v_i = \sqrt{\frac{v_i}{\|V\|_1}} \qquad i = 1, 2, ..., 128 \tag{4.3}$$

K-means clustering is performed on the rootSIFT feature from the training set to form a codebook, where the size of the cookbook is denoted by M. A sequence of grids divides the cell image into 4^l finer spatial regions of each scale, where $l \in \mathbb{R}^L$ denotes the level of resolution. We weight each of histogram according to Mercer kernels [118], and then concatenate those histogram into one descriptor of dimensionality

$$D = M \sum_{l=0}^{L} 4^l. \tag{4.4}$$

4.2.4 Classifier

The object-graph based feature and the spatial pyramid matching histogram are then fused into one representative vector. We normalize the combined feature with L_2 norm before training. Random forest with CART is proposed to train a model for classifying the stained pattern.

4.3 Experiments

4.3.1 Dataset

I3A (Indirect Immunofluorescence Image Analysis)-2012 dataset that introduced in ICPR HEp-2 cell classification contest dataset was used to evaluate our method. The dataset contains 28 slide images which captured by a 50 W mercury-vapor lamp and digitalized with a resolution of 1388×1038. Pathologists categorized the dataset according to different patterns and fluorescence intensities (positive or intermediate). The slide image was partitioned into a training set of 721 cell images and a test set of 734 cell images. Table 4.1 shows the detailed information of I3A dataset. The top row of each cell type reports the positive stained information, while bottom row for intermediate. Each item gives the number of slide image and cell image (in parentheses).

Type	Training Set	Test Set	Total
homogrneous	2(103) 1(47)	1(119) 1(61)	3(222) 2(108)
Coarse Speckled	1(68) 1(41)	2(68) 1(33)	3(136) 2(74)
Centromere	1(89) 2(119)	2(84) 1(65)	3(165) 3(184)
Nucleolar	1(56) 1(46)	1(73) 1(66)	2(129) 2(112)
Fine Speckled	1(46) 1(48)	1(51) 1(63)	2(97) 2(111)
Cytoplasmic	1(34) 1(24)	0(0) 2(51)	1(36) 3(75)
Total	7(396) 7(325)	7(395) 7(339)	14(791) 14(664)

Table 4.1: Composition of the HEp-2 I3A-2012 images dataset.

4.3.2 Results

All experiments strictly followed the rule of contest. We resized the cell image to 128×128, and expanded the training data by rotating the resized image 90°, 180° and 270°. The random forest classifier contains 200 decision trees.

We first tested the performance of the object-graph algorithm with different numbers of background and foreground thresholds. I were chosen from 3 to 5, while J were from 4 to 8. Table 4.2 shows the relationship between the accuracy and the number of thresholds. The highest accuracy is 66.45% with 5 background thresholds and 5 foreground thresholds respectively.

Accuracy (%) \ J / **I**	**5**	**6**	**7**	**8**
3	59.68	58.36	55.90	54.11
4	62.45	61.79	60.95	58.64
5	66.45	64.16	62.35	59.67

Table 4.2: The variance of classification accuracy against the number of binary mask thresholds.

The level of the pyramid in our experiment is $L = 2$, and the dictionary size is $M = 500$, which gives the best classification rate of 69.18%. Thus, the final SPM feature has 10500 dimensions. There is no significant increase in accuracy beyond $M = 500$. Further, the computational complexity increased dramatically with dictionary size.

The object-graph based feature and SPM were finally combined as a 10577-dimensional feature vector. Table 4.3 shows the confusion matrix of the combined feature. Our method gives the best accuracy on centromere cells followed by homogeneous. Fine speckled cells were the main source of error, which were easily misclassified as homogeneous. Table 4.4 compares the overall classification rate with other methods. The SPM has a better performance than RICLBP, the winner of ICPR 2012 contest. The overall accuracy of our method is comparable to the winner of the ICPR 2014 contest.

	H	**CS**	**C**	**N**	**FS**	**Cy**
H	162	6	0	0	12	0
CS	5	79	4	0	13	0
C	0	1	136	12	0	0
N	3	14	10	83	4	25
FS	49	27	13	0	25	0
Cy	0	2	2	0	0	47

Table 4.3: The cell level confusion matrix of our algorithm.

Method	Overall Accuracy (%)
Nosaka [62]	68.53
Kazanov [121]	70.57
Shen + BOW [71]	74.39
Object-Graph	66.45
SPM	69.18
Object-Graph+SPM	72.47

Table 4.4: Performances of different algorithms on the I3A-2012 dataset.

4.4 Summary

In this chapter, we presented a novel automatic system for HEp-2 cell image classification. A set of morphological properties are extracted from a set of circle-transformed images. Spatial pyramid matching is introduced to describe the texture information. Those two features are fused into one vector, and the random forest is utilized to classify the IIF stained patterns. The proposed scheme was evaluated on I3A-2012 HEp-2 dataset and achieved competitive accuracy compared to other hand-craft feature based classification methods.

There exists some limitations in our model. For instance, the SPM feature is not rotation invariance. Thus, future work will try to design new BoW model to strengthen the robustness of image rotation. Another is that two different features are added directly without pooling. Whether pooling or feature dimension reduction for this model is necessary will be examined. In addition, more classifiers such as SVM will be used in the future work.

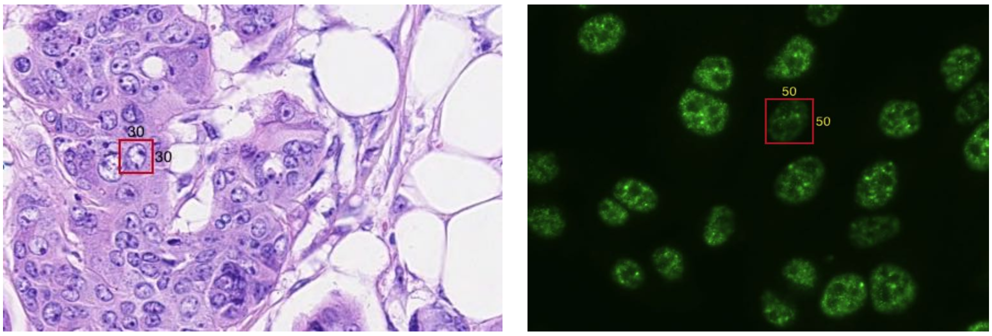
Chapter 5

Cell Classification via Deep Autoencoding-Classification Network

In the last chapter, we have studied the cell classification problem by object-graph based feature. Nevertheless, it is proposed especially for HEp-2 cell and lack of robustness. Motivated by the significant performance and robustness of deep learning on various computer vision tasks [78], we present a novel deep learning model termed Deep Autoencoding-Classification Network (DACN) for cell classification. Our proposed model consists of an autoencoder and a normal classification convolutional neural network (CNN), while the two architectures share the same encoding pipeline. We evaluate the proposed model on three publicly available datasets. In a series of experiments, we show that this architecture is able to achieve higher cell classification accuracy than CNN, and is particularly effective when the training dataset is small which is often the case in medical imaging applications.

5.1 Introduction

Automatic cell classification approaches on pathological digital images are constrained by several factors. The dimension of human cells is between 1 to 100 micrometres, which means the corresponding image resolution is relatively low (image size normally is smaller than 30×30 pixels with $20\times$ optical magnification). It is a challenging problem to extract information from such a small image. On the other hand, high variability in the appearance of the same type of cells, overlapping, and heterogeneity also affect the classification accuracy as shown in Figure.5.1. Training CAD systems need manually label digital image, however, these problems make labelling work extremely laborious. Thus, maximizing the utilization training set, which extracting so much useful information as possible is the key for cell classification CADs.



(a) H&E (b) IIF

Figure 5.1: Examples of digital histopathology cell images. (a) is H&E stained colorectal adenocarcinomas image with $20\times$ optical magnification; (b) is $40\times$ magnification IIF HEp-2 cell image.

Recent years have witnessed the increasing popularity of deep learning methods in computer vision applications. Deep convolutional networks have significantly promoted the image classification performance [79, 80, 81]. However, different from large scale nature image visual recognition task, the CNN model used for histopathological images are much simpler because of the small training set and image size [86, 84].

Since previous works have proved that unsupervised learning is extremely useful in deep neural network [122, 123], in this chapter, we propose Deep Autoencoding-Classification Network (DACN), which maximize the utilization of the limited training set using unsupervised learning. Inspired by [123, 124], DACN embeds a deep convolutional autoencoder into the normal label prediction network as shown in Figure 5.2, the two networks share the encoding pipeline. The DACN model is jointly optimized according to two loss functions, one is the classification error and the other is image reconstruction error. The autoencoding learning procedure forces the encoder output to capture information which will enable to reconstruct the input. This ensures that the encoder part captures the full information of the input to improve performances in the classification network. The proposed method is evaluated on three datasets: I3A-2012 HEp-2 image dataset, I3A-2013 HEp-2 image dataset, and CRCHistoPhenotypes dataset. We will show that this architecture is particularly effective when training dataset is small and is able to mitigate the overfitting problem. We will present experimental results to demonstrate the effectiveness of our new model.

5.2 Deep Autoencoding-Classification Network

We begin with a review of the traditional CNN model. A convolutional classification network F is a composition of a sequence of layers that maps image data of the input space $\mathscr{X}$ to output label space $\mathscr{Y}$ ($\mathsf{F} : \mathscr{X} \rightarrow \mathscr{Y}$). Given an input $x \in \mathscr{X}$ with its label $y \in \mathscr{Y}$, CNN is trained by solving the optimization problem:

$$\arg\min_{\mathbf{W}} \frac{1}{N} \sum_{i}^{N} l(\mathsf{F}(x_i, \mathbf{W}), y_i) \tag{5.1}$$

where l is the cross-entropy loss, and $\mathbf{W}$ is the parameters of the classification model.

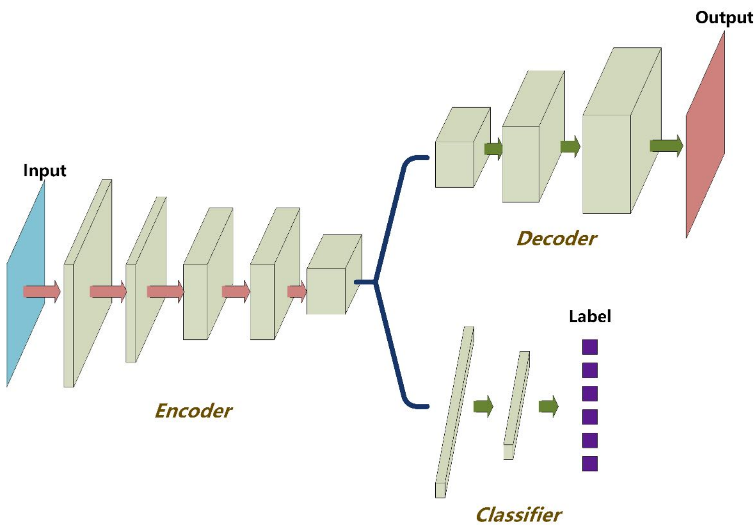


Figure 5.2: The overall architecture of deep convolutional autoencoding-classification network.

In our DACN model, we separate the CNN pipeline into an *Encoder* $\mathsf{f_{en}}$ and *Classifier* $\mathsf{f_{cl}}$, and add an additional convolutional *Decoder* $\mathsf{f_{de}}$ after $\mathsf{f_{en}}$ as shown in Figure 5.2. Therefore, the DACN model is composed of an unsupervised deep autoencoder $\mathsf{F_A}$ and a supervised classification model $\mathsf{F_C}$.

$$\mathsf{F_A} = \mathsf{f_{en}} \circ \mathsf{f_{de}}, \tag{5.2}$$

$$\mathsf{F_C} = \mathsf{f_{en}} \circ \mathsf{f_{cl}}. \tag{5.3}$$

The parameters of DACN are also divided for two sub-models according to the architecture: $\mathbf{W}_A = \{\mathbf{W}_{en}, \mathbf{W}_{de}\}$ and $\mathbf{W}_C = \{\mathbf{W}_{en}, \mathbf{W}_{cl}\}$. Two sub-models share the encoder part. The rationale of DACN is that the encoder trained by autoencoder can provide useful information for classifier. Because it forces the hidden layer outputs to contain input information, which will enable a reconstruction of the input.

The model is optimized using multi-task learning according to two learning objectives. The deep autoencoder is trained to minimize the error between input image x_j and the reconstructed image $\mathsf{F}_\mathsf{A}(x_j)$ as:

$$\mathscr{L}_A(\{\mathbf{W}_{en},\mathbf{W}_{de}\}) = \sum_j^M l_A(\mathsf{F}_\mathsf{A}(x_j,\{\mathbf{W}_{en},\mathbf{W}_{de}\}),x_j), \tag{5.4}$$

where l_A forms the traditional square error $\|\mathbf{x}-\mathsf{F}_\mathsf{A}(x_j)\|^2$. While the loss function of classifier part is similar to Eq.(1):

$$\mathscr{L}_C(\{\mathbf{W}_{en},\mathbf{W}_{cl}\}) = \sum_i^N l_C(\mathsf{F}_\mathsf{C}(x_i,\{\mathbf{W}_{en},\mathbf{W}_{cl}\}),y_i). \tag{5.5}$$

Consequentially, the total loss of CACN model is

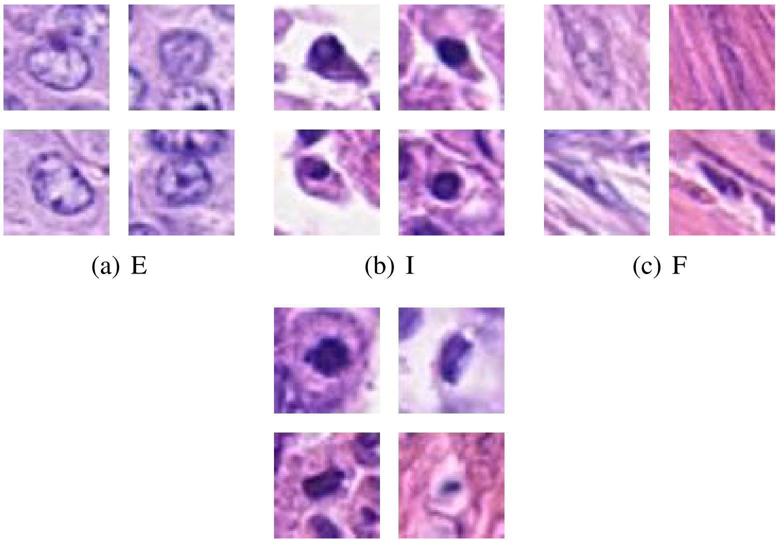
$$\mathscr{L}_{tot} = \lambda\mathscr{L}_A(\{\mathbf{W}_{en},\mathbf{W}_{de}\}) + \mathscr{L}_C(\{\mathbf{W}_{en},\mathbf{W}_{cl}\}). \tag{5.6}$$

where λ is the trade-off between two different losses. Similar to [124], the deep autoencoder and CNN classifier are trained alternately.

5.3 Experiments

5.3.1 Datasets

In this work, the proposed deep learning model is evaluated on three different datasets with different stain types and magnifications (e.g., CRC, I3A-2012, and I3A-2013). For CRC and I3A-2013, datasets were randomly divided into three subsets: 64% for training, 16% for validation, and 20% for the test; while I3A-2012 has specified training and test set. Since the size of I3A-2012 is small, the proposed model is evaluated on CRC and I3A-2013 datasets without I3A-2012.



(a) E (b) I (c) F

(d) M

Figure 5.3: Example patches of different types of nuclei in the CRC dataset. (a): *Epithelial*, (b): *Inflammatory*, (c):*Fibroblast*, and (d): *Miscellaneous*.

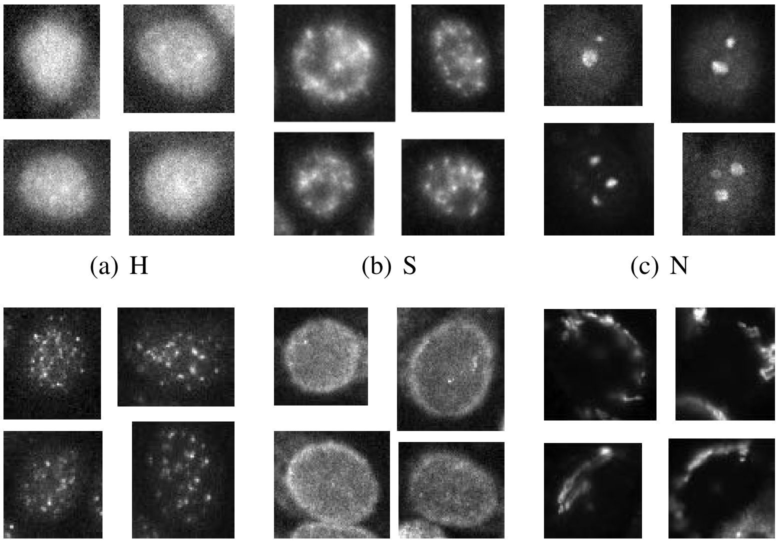
CRCHistoPhenotypes Dataset

The CRC dataset [84] contains 100 H&E stained digital images of colorectal adenocarcinomas with a common size of 500×500 pixels. The images were cropped 10 whole slide images (WSI) from 9 patients with no overlapping area, and scanned by Omnyx VL120 scanner at the high resolution of $20\times$ optical magnification. The cropped areas contain both normal and malignant regions, and also comprises over-staining and out-of-focus areas. There are 20,619 nuclei that have been labelled with associated classes, i.e. *Epithelial* (E), *Inflammatory* (I), *Fibroblast* (F), and *Miscellaneous* (M) (see Figure 5.3). The nuclei labelled with inflammatory include lymphocyte plasma, neutrophil and eosinophil; while miscellaneous nuclei consist of adipocyte, endothelium, mitotic figure, nucleus of necrotic. Since the cell images are cropped from 500×500 images with the window size of 27×27, we eliminate the cell at the edge of the image that only uses 20,436 cell images.

I3A-2012 Dataset

The dataset is same as that used in Chapter 4, which contains six different patterns of *Homogeneous* (H), *Coarse Speckled* (CS), *Centromere* (C), *Nucleolar* (N), *Fine Speckled* (FS), and *Cytoplasmic* (Cy). The example images can be seen in Figure 4.1.

I3A-2013 Dataset



(a) H (b) S (c) N

(d) C (e) NM (f) G

Figure 5.4: Example images of different types of nuclei in the I3A-2013 dataset. (a-f) are *Homogeneous*, *Speckled*, *Nucleolar*, *Centromere*, *Nuclear Membrane* and *Golgi* respectively.

We also evaluate the proposed method on the I3A-2013 training dataset which used in ICIP 2013 and ICPR 2014 HEp-2 cell classification contests. The dataset contains 13, 596 images, which acquired through fluorescence microscope with 40-fold magnification. Different from the I3A-2012 HEp-2 cell contest dataset, each image belongs to one of the six staining patterns: *Homogeneous* (H), *Speckled* (S), *Nucleolar* (N), *Centromere* (C), *Nuclear Membrane* (NM) and *Golgi* (G) as shown in Figure 5.4.

5.3.2 Image preprocessing and data augmentation

The cropped cell image size from CRCHistoPhenotypes dataset is 27×27, and then resize to 56×56; while all HEp-2 cell images are resized to 78×78. The CRCHistoPhenotypes uses all three RGB channels; while only the green channel is extracted for IIF stained HEp-2 image, since most information is in the green channel.

Since intermediate HEp-2 cell images have intermediate staining with extremely low intensity, we first pre-process all images in order to extract reliable information for classification. Specifically, all images are normalized by following gray level stretch equation:

$$I_{new} = \frac{I_{ori} - \min(I_{ori})}{\max(I_{ori}) - \min(I_{ori})} \tag{5.7}$$

where I_{ori} denotes the input image, and $\min(\cdot)$ and $\max(\cdot)$ are computing the minimum and maximum intensity value respectively.

The data balance and the number of training examples directly affect the performance of deep learning networks. The normal method for cell classification work is rotating each image with fixed angle steps [85, 87]. In this work, we first balance the data by mirror horizontally and rotation the image with fixed angles $(90°, 180°, 270°)$. During CNN training, the training set is augmented in three ways at the same time: rotation with random angles, mirror horizontally, and randomly shift the image horizontally and vertically within 10%.

5.3.3 Network architecture

Architectures

The CNN architecture utilized for two different datasets are same. The encoder parts are based on AlexNet [79], and the proposed encoder architectures are more elegant and efficient. The input image size are $56 \times 56 \times 3$ and $72 \times 72 \times 1$ for CRC and I3A datasets respectively. While the corresponding reconstructed output is the half size of the input, and the reconstruction image for CRC image is grayscale. The reason is that only the morphology of cells is useful for classification instead of detail texture information. Thus, reconstructing a half size image would force the model to learn more morphological information for reconstruction.

The encoding path uses 2×2 maxpooling operation for downsampling; while the expansive path upsamples previous feature channels by 2×2. The detailed architectures of the convolutional encoder and decoder are shown in Table.5.1, while f_{cl} contains two fully connected (FC) layers before softmax. To avoid overfitting, we add dropout layers with 0.1 after each convolutional block in encoding pipeline, and two dropout layers after the last maxpooling layer and FC layer with 0.3 and 0.5 respectively.

5.3.4 Implementation details

The model is trained by alternately minimizing $\mathscr{L}_A$ and $\mathscr{L}_C$. In the implementation, the decoder uses sigmoid for reconstruction, and is configured with optimizer of Adadelta [125]; while stochastic gradient descent (SGD) is used for softmax.

The unsupervised autoencoder is firstly trained before classification as a regularizer for small training set (In this project, if the training set is smaller than

Layer	Encoder f_{en}		Decoder f_{de}	
	Type	Input/Filter Dimensions	Type	Input/Filter Dimensions
0	I	56×56×1		
1	C	16×7×7	U	2×2
2	M	2×2	C	64×3×3
3	C	32×5×5	U	2×2
4	C	32×3×3	C	32×3×3
5	M	2×2	C	1×1×1
6	C	48×3×3		
7	C	48×3×3		
8	C	48×3×3		
9	M	2×2		
10	C	64×3×3		
11	C	64×3×3		
12	C	64×3×3		

Table 5.1: Architecture of the encoder and decoder parts of the Deep Autoencoding-Classification Network. The pipeline consists of Input (I), Convolutional layer (C), Maxpooling layer (M), and Upsampling layer (U).

test set). Autoencoding as pretraining was first proposed in [123], which aims to improve the supervised deep learning classification accuracy. Although this method has fallen out of fashion, our experimental result shows that the unsupervised training helps to provide a better starting point for the classifier network than random initialization on a small training set.

The stopping criterion during training is determined by monitoring the two losses. The process of training autoencoder is stopped when the validation reconstruction loss converges; while classification process is stopped after the training loss stabilizes or monitoring overfitting. It is empirical to choosing the values of λ during the alternative training procedure. In this work, the weighting factor λ is initially set to a range between 2 and 4 for autoencoder. During alternative training procedure, we decrease the value to [0.6 - 1]. Moreover, learning rate decay is utilized to increase performance, which decreases the learning rate values of both processes after a loop. In this work, the total training process needs 3-4 loops.

We implement our model using Keras based on Tensorflow [126]. The network was trained on single GeForce GTX 1080ti GPU with 11GB RAM.

5.3.5 Performance analysis

Model Evaluation

The proposed DACN contains two outputs, and each has its own loss function. In this work, we propose alternatively optimize the DACN model, while the model can also be optimized at the same time with one combination loss. During training procedure, λ is the key parameter in DACN model, and λ is decreased during the training procedure of proposed framework. Here, we first conduct an experiment to evaluate the effects of different training modes (e.g., alternative or one-step) and different λ trends on DACN.

We have trained a total of six different models: DACN-O-INC, DACN-O-INV, DACN-O-DEC, DACN-A-INC, DACN-A-INV, and DACN-A-DEC. 'A' and 'O' mean alternative learning and one-step learning respectively; while 'INC', 'INV', and 'DEC' are the trend of λ of increasing, invariant, and decreasing respectively. Our proposed approach is DACN-A-DEC.

Method	Overall Accuracy (%)	
	CRC	I3A-2013
DACN-O-INC	78.63	91.96
DACN-O-INV	79.40	93.77
DACN-O-DEC	80.34	94.04
DACN-A-INC	82.58	98.49
DACN-A-INV	83.43	98.53
DACN-A-DEC	**83.95**	**98.79**

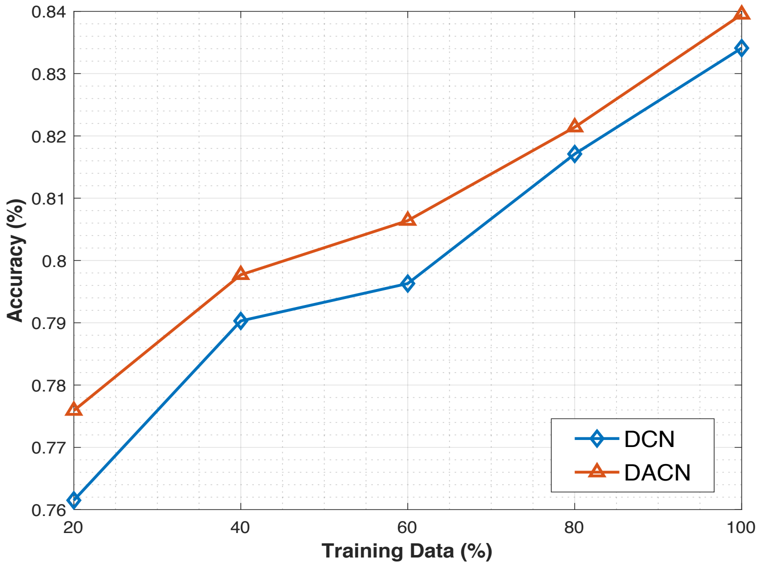
Table 5.2: Comparing classification accuracy of different settings of DACN on CRC and I3A datasets.

Table.5.2 compares the results of different models on both two datasets. It can be seen that the performances of alternatively training significantly outperform the approach of jointly optimizing one combination loss on both datasets, which means optimizing one combination loss is not suitable for this task. The models with decreasing λ achieve better results followed by invariant λ. Thus, acting as a regularizer, the autoencoding procedure could be set with a relatively high learning rate. However, increasing λ or keeping λ invariant would affect the classification performance.

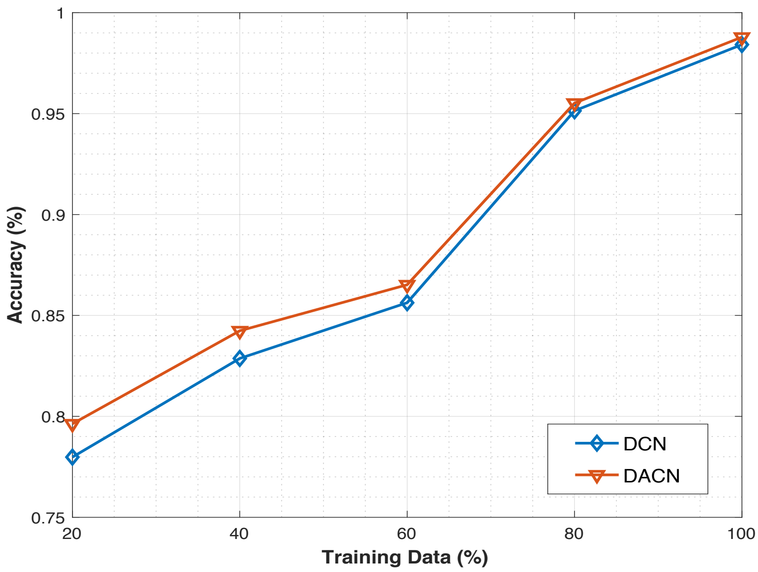
Evaluation of training size

One of our intuitions is that when the training dataset is small, our model should have more advantage over the traditional deep convolutional network (DCN). To test this, we train the DACN model using different sizes of training data and compare the results with DCN without the autoencoding component.

Figure 5.5 shows the classification accuracy on two datasets when 20%, 40%, 60%, 80% and the full training set are used to train the networks. As can be seen, for a smaller training set, the DACN outperforms DCN by a large margin, and in all cases, DACN has a higher performance gain over DCN. One possible explanation for this could be that the encoding procedure ensures the information fed to the classification network captures the most important information of the input (because it enables a reconstruction of the input). Another possible explanation is that the autoencoding procedure acts as a regularizer to ensure the encoder representing the input in a way that facilitates classification.



(a) CRC dataset



(b) I3A-2013 dataset

Figure 5.5: A comparison of DCN and DACN results when the training data size varies. For smaller training dataset, DACN achieves higher performance gain while in all cases the new method outperforms traditional model.

5.3.6 Results

CRC Dataset

As a new released dataset, few cell classification model test on CRCHistoPhenotypes images. Table 5.3 compares our proposed model with the baseline CNN model introduced in [84], and DCN that using same encoding pipeline as DACN while without autoencoder.

Method	Overall Accuracy (%)
Sirinukunwattana *et al.*	71.20
DCN	83.41
DACN	**83.95**

Table 5.3: Performances of different algorithms on the CRC dataset.

It can be seen that both DCN and DACN surpass the baseline CNN model with a large margin. The reason is that the baseline model only contains two convolutional layers. Thus, network depth enables the models to get better feature and accuracy for cell classification.

The confusion table of proposed method is presented in Table 5.4. The DACN model gives the worst recognition accuracy on *Miscellaneous* class. There are two main reasons: *Miscellaneous* is a mixed class with four different kinds of cells; and it has the minimum image number, which only contains about a third of images in *Epithelial* class. The accuracy will increase to 85.94% by eliminating the *miscellaneous* class.

	E	**I**	**F**	**M**
E	1298	75	36	15
I	39	825	107	54
F	18	121	1065	68
M	23	44	58	241

Table 5.4: Confusion table for DACN of CRC dataset.

I3A-2012 Dataset

Since I3A-2012 only have 721 training images and 734 test images, we employed a simpler CNN model with six convolutional layers, which get rid of layer 8 and 12 (see Table 5.1) in encoding pipeline. The detailed architecture can be seen in [127].

	Method	Overall Accuracy (%)
Hand-Craft	Nosaka [62]	68.53
	Object-Graph+SPM	72.47
	Shen [71]	74.39
Deep Learning	Gao [86]	74.8
	Phan [87]	77.1
	DCN	76.68
	DACN	**78.21**

Table 5.5: Performances of different algorithms on the I3A-2013 testing set.

	H	**CS**	**C**	**N**	**FS**	**Cy**
H	151	1	1	1	24	0
CS	2	77	3	3	17	2
C	1	11	121	11	0	5
N	11	0	1	114	2	12
FS	28	11	3	3	66	3
Cy	0	3	0	1	0	45

Table 5.6: Confusion table for DACN of I3A-2012 dataset.

Table 5.5 compares the classification performance of our method with state-of-the-art algorithms on the I3A-2012 test set. DCN and DACN achieve better performances than Gao [86] and Phan [87], which means that the proposed deep CNN model is more advanced. Our proposed algorithm surpasses all state of the art methods, which achieves 78.21% accuracy rate by training on the I3A-2012 training set. The confusion table for DACN is illustrated in Table 5.6. Similar to the result on CRC dataset, the model gives the worst classification accuracy on the *Cytoplasmic* class with the minimum image number.

I3A-2013 Dataset

The proposed DACN model is compared with hand-craft based methods, such as the Bag-of-features (BoF) model [128], the Fisher Vector (FV) model [128]; and recently introduced CNN based model, i.e, Gao *et al* [128] proposed model with 5 layers and Li *et al* [129] proposed model with complex residual inception. The DCN without autoencoding procedure is also involved for comparison. Table 5.7 shows the average classification accuracy of different methods.

	Method	Overall Accuracy (%)
Hand-Craft	BoF	94.38
	FV	96.07
Deep Learning	Gao [128]	97.24
	Li [129]	98.49
	DCN	98.42
	DACN	**98.79**

Table 5.7: Performances of different algorithms on the I3A-2012 dataset.

All deep learning based approaches outperform traditional hand-craft based methods, which means that CNN models are able to extract better features. Our proposed algorithm surpasses all state-of-the-art methods, which achieves 98.79% accuracy rate. The results demonstrate that increasing network complexity is not the only approach to learn better feature representations; our proposed model with the autoencoding pipeline is able to extract more important feature for cell classification. Table 5.8 shows the confusion matrix of proposed methods on I3A-2013. It can be seen that the performance on six different classes is relatively even.

5.4 Summary

In this chapter, we have presented a novel deep learning system for cell image classification. The model has the same spirit of multi-task learning, which

	H	**S**	**N**	**C**	**NM**	**G**
H	498	2	0	0	0	2
S	6	543	5	1	3	2
N	0	0	520	1	1	0
C	0	3	1	541	0	1
NM	1	0	0	0	442	0
G	0	0	3	0	1	143

Table 5.8: Confusion table for DACN of I3A-2013 dataset.

combines a convolutional autoencoder and a traditional classification CNN. The autoencoder is proposed to learn additional pattern information for the classification task, and also acts as a regularizer for a small training set. The reconstructed images from the autoencoder resemble the input images, which indicates that two models can be successfully merged for effectively learning cell pattern information. The proposed model is evaluated on two datasets of different stains. The experimental results show that our new DACN method provides a considerable improvement over the traditional CNN models, and outperforms all known state-of-the-art approaches on the three datasets.

Our method has some limitations. First, it is difficult to determine a series of fixed values of λ during training and the number of training loops, since those parameters are chosen according to the losses of encoding and classification during the training process. Future works will try to overcome this limitation. In addition, recent proposed residual network architecture is not used in our proposed model. Future works will try to utilize more advanced architectures to promote the performance.

Chapter 6

Automatic Histochemical Scoring in Tissue MicroArray of Breast Cancer

In this chapter, we present an end-to-end deep learning system which directly predicts the H-Score automatically. The innovative characteristics of our method are that it is inspired by the H-Scoring process of the pathologists where they count the total number of cells, the number of tumour cells, and categorise the cells based on the intensity of their positive stains. Our system imitates the pathologists' decision process and uses one fully convolutional network (FCN) to extract all nuclei region (tumour and non-tumour), a second FCN to extract tumour nuclei region, and a multi-column convolutional neural network which takes the outputs of the first two FCNs and the stain intensity description image as input and acts as the high-level decision making mechanism to directly output the H-Score of the input TMA image. In additional to developing the deep learning framework, we also present methods for constructing positive stain intensity description image and for handling discrete scores with numerical gaps.

Whilst deep learning has been widely applied in digital pathology image analysis, as far as the present author's knowledge extends, this is the first end-to-end system that takes a TMA image as input and directly outputs a clinical score. We will present experimental results which demonstrate that the H-Scores predicted by our model have very high and statistically significant correlation with experienced pathologists' scores and that the H-Scoring discrepancy between our algorithm and the pathologist is on par with that between the pathologists. Although it is still a long way from clinical use, this work demonstrates the possibility of using deep learning techniques to automatically and directly predicting the clinical scores of digital pathology images.

6.1 Introduction

Breast cancer (BC) is a heterogeneous group of tumours with varied genotype and phenotype features [110]. Recent research of Gene Expression Profiling (GEP) suggests that BC can be divided into distinct molecular tumour groups [130, 131]. Personalised BC management often utilizes robust commonplace technology such as immunohistochemistry (IHC) for a tumour molecular profiling [132, 133].

Diaminobenzidine (DAB) based IHC techniques stain the target antigens (detected by biomarkers) with brown colouration (positive) against a blue colouration (negative) counter-stained by Hematoxylin (see Figure .6.1 for some example images). To determine the biological class of a tumour, pathologists will mark the nuclei activity biomarkers through a microscope and give a score based on a semi-quantitative assessment method called the modified histochemical scoring (H-Score) [31, 134]. The H-Scores of tissue samples stained with different biomarkers is combined together to determine the biological class of a case. Clinical decision making is to choose an appropriate treatment from a num-

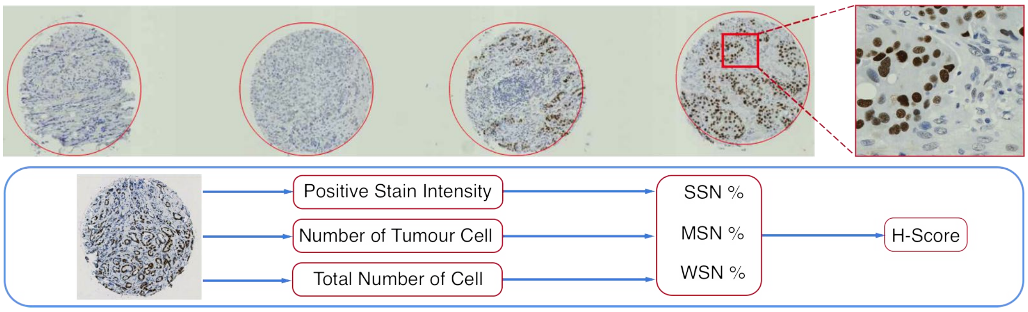


Figure 6.1: Top: Example images extracted from digital TMA slides. Each red circle contains one TMA core stained by Diaminobenzidine-Hematoxylin (DAB-H). The brown colours indicate positive and the blue colours indicate negative. Bottom: A schematic illustration of the traditional manual H-Scoring procedure: It needs to first count the total number of nuclei, then the number of strongly stained, moderately stained and weakly stained tumour nuclei, respectively. The final H-Score is then calculated according Eq.6.1.

ber of available treatment options according to the biological class of the tumour. For instance, one of the methods for stratifying different molecular classes is the Nottingham Prognosis Index Plus (NPI +)[110] which uses 10 breast cancer relevant biomarkers to stain tumour tissues prepared on tissue microarray (TMA). Tissue samples stained by each of these 10 biomarkers are given a histochemical score (H-Score) and these 10 scores together will determine the biological class of the case.

Therefore, H-Score is one of the most important pieces of information for molecular tumour classification. When the tumour region occupies more than 15% of the TMA section, an H-Score is calculated based on a linear combination of the percentage of strongly stained nuclei (SSN), the percentage of moderately stained nuclei (MSN) and the percentage of weakly stained nuclei (WSN) according to:

$$\mathrm{H-Score} = 1 \times \mathrm{WSN} + 2 \times \mathrm{MSN} + 3 \times \mathrm{SSN} \tag{6.1}$$

The final score has a numerical value ranges from 0 to 300. Thus, the histochemical assessment of the TMA's is based on the following semi-quantitative information: the total number of cells, the number of tumour cells and the stain intensity distributions within the tumour cells. In clinical practice, diagnosis requires averaging two experienced pathologists' assessments. Manually marking the positively stained nuclei is obviously a time consuming process. As visual assessment of the TMA's is subjective, there is the problem of inter-observer discrepancy and the issue of repeatability. The semi-quantitative nature of the method (strongly stained, moderately stained and weakly stained, the definitions of strong, moderate and weak cannot be precise and subjective), makes it even more difficult to ensure inter-subject as well as intra-subject consistency.

With the increasing application of clinicopathologic prognosis, Computer Aided Diagnosis (CAD) systems have been proposed to support the pathologists' decision making. The key parameters in tissue image assessment include the number of tumour cells, the positive staining intensities within these cells and the total number of all cells in the image. To classify the positively stained pixels and their stain intensity, methods such as colour deconvolution that perform mathematical transformation of the RGB image [51] [1] are widely used to separate positive stains from negative stains. Numerous computer-assisted approaches have been proposed for cell or nuclei detection and segmentation [22]. Most literature on histopathology image analysis perform various low-level quantification steps, there is still little attempt to perform an end-to-end assessment of the image directly.

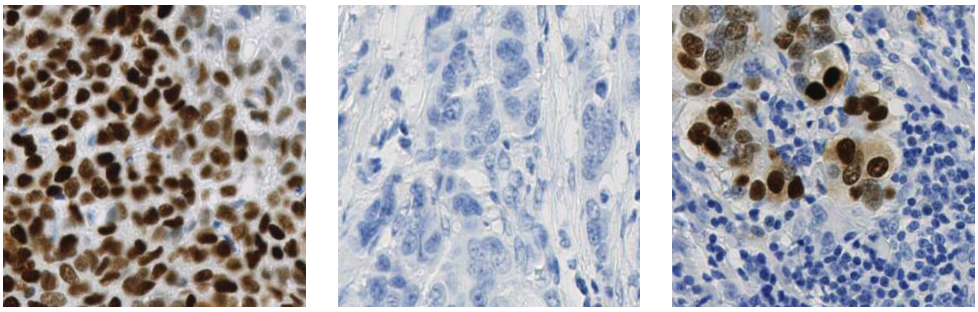
Is it possible to develop a CAD model that would directly give a high-level assessment of a digital pathological image, just like an experienced pathologist would, for example, to give out an H-Score directly? In an attempt to answer this question, we propose an end-to-end deep learning system for directly predicting the H-Scores of breast cancer TMA images, see Figure 6.6. Instead of pushing the raw digital images into the neural network directly, we follow a similar process that pathologists use for H-Score estimation. We first construct a stain intensity nuclei image (SINI) which only contains nuclei pixels and their corresponding stain intensity information, and a stain intensity tumour image (SITI) which only contains tumour nuclei pixels and their corresponding stain intensity information. The SINI and SITI block irrelevant background pixels while only retain useful information for calculating the H-Score. These two H-Score relevant images are then fed into a dual-channel convolutional neural network with two input pipelines, which are finally merged into one pipeline to give an output (H-Score). To the best of our knowledge, this is a first work that attempts to develop deep learning based TMA processing model that directly outputs the histochemical scores. We will present experimental results which demonstrate that the H-Scores predicted by our model have a high and statistically significant correlation with experienced pathologists' scores and that the H-Scoring discrepancy between our algorithm and the pathologists is on par with that between the pathologists. Although it is still perhaps a long way from clinical use, this work nevertheless demonstrates the possibility of automatically scoring cancer TMA's based on deep learning.

6.2 Problem and Methodology

An immunohistochemical assessment can be formulated as a model F that maps the input images from the input space $\mathscr{I}$ to the a label space $\mathscr{L}$. Given an input image $I \in \mathscr{I}$, its label $l \in \mathscr{L}$ is assigned according to the quantitative information of positive staining intensity P_s, the number of tumour cells N_t and total number of cells N_e in the image x:

$$l = \mathsf{F}(I|P_s, N_t, N_e), \tag{6.2}$$

Traditional assessment methods have at least three unsolved issues for both the



(a) (b) (c)

Figure 6.2: Examples of challenging cases of quantitative measurement of biomarkers based on visual assessment. (a) A variety of stain intensities; (b) unclear staining and overlapping of nucleus; (c) Size differences between different type of nucleus.

pathologists and the CAD systems. Firstly, the positive staining intensity needs to be categorized into four classes: *unstained*, *weak*, *moderate*, and *strong*. However, there is no standard quantitative criterion for classifying the DAB stain intensity. Thus, two pathologists often classify the same staining intensity into two different categories or two different intensities into the same category. Furthermore, the human visual system may pay more attention to strongly stained regions but they are often surrounded by a variety of staining intensities [4], which may also affect the assessment results. Secondly, cell/nuclei instance counting is a very important parameter in the assessment. Nevertheless, both human and computer still cannot deal with the difficulty of counting overlapping cells very well. Moreover, variability in the appearance of different types of nucleus, heterogeneous staining, and the complex tissue architectures make individually segmenting cell/nuclei a very challenging problem. Thirdly, the apparent size differences between tumour nuclei and normal nuclei will affect the quantitative judgement of tumour nuclei assessment. Examples of these challenging cases are illustrated in Figure 6.2.

To tackle the problem mentioned above, we propose to develop a convolutional neural network (CNN) based CAD framework for biomarker assessment of TMA

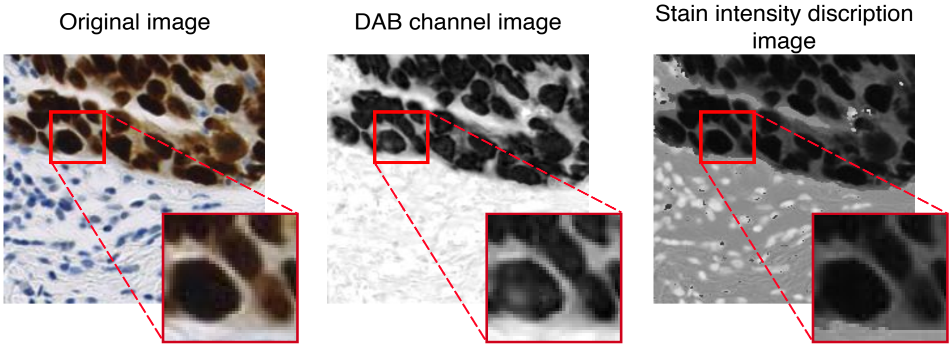


Figure 6.3: A comparison of different images generated during the process of stain intensity description. The highlighted subimage contains strongly stained nuclei.

images. Instead of using CNN as a feature extractor or for low level processing such as cell segmentation only, we have developed an end-to-end system which directly predicts the biomarker score (H-Score). The innovative characteristic of our method is that it is inspired by the H-Scoring process of the pathologists where they count the total number of nuclei and the number of tumour nuclei and categorise tumour nucleus based on the intensity of their positive stains. In the complete system, as illustrated in Figure 6.6, one fully convolutional network (FCN) is used to extract all nuclei region which acts as the step of counting all nucleus and capture all foreground information, another FCN is used to extract tumour nuclei region which acts as the step of counting all tumour nucleus. To mimic the process of categorising tumour nuclei based on their positive stain intensities, we derive a stain intensity image which together with the outputs of the two FCNs are presented to another deep learning network which acts as the high-level decision-making mechanism to directly output the H-Score of the input TMA image.

6.2.1 Stain intensity description

Although various DAB stain separation methods have been proposed [1, 2], few work studied the stain intensity description and grouping. Since there is no formal definitions for the boundaries between stain intensity groups (e.g, *strong*, *moderate*, *weak*), previous works used manually defined thresholds for pixel-wise classification to segment positive stains into each stain group [4].

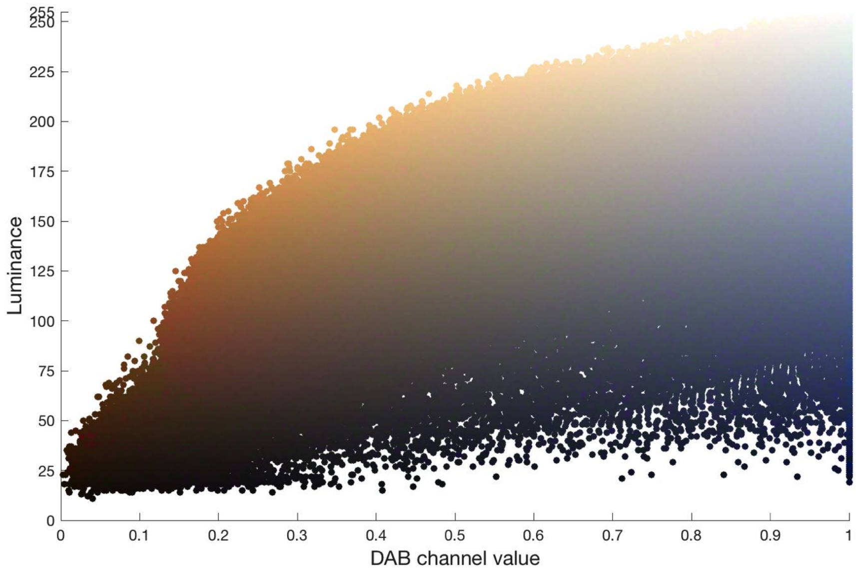


Figure 6.4: Visualization of pixel colours of the DAB-H images along the luminance axis and the colour deconvolution DAB axis.

In this work, we propose to directly use the luminance values of the image to describe the staining intensity instead of setting artificial intensity category boundaries. The original RGB image I is first transformed into three-channel stain component image ($I_{DAB-H} = [I_{DAB}, I_H, I_{Other}]$) using colour deconvolution [1]:

$$I_{DAB-H} = M^{-1} I_{OD}, \tag{6.3}$$

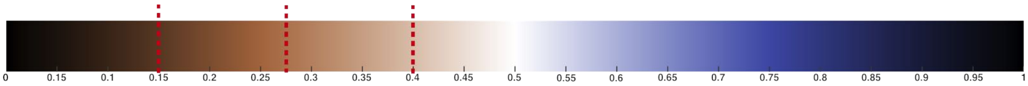


Figure 6.5: An illustration of the value of l_{la} and its corresponding stain intensity. The red dot lines are the thresholds of stain intensity groups [4].

where M is the stain matrix composed of staining colours equal to

$$\begin{bmatrix} 0.268 & 0.570 & 0.776 \\ 0.650 & 0.704 & 0.286 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \tag{6.4}$$

for DAB-H stained images, and I_{OD} is Optical Density converted image calculated according Lambert-Beers law:

$$I_{OD} = -log(\frac{I}{I_0}), \tag{6.5}$$

$I_0 = [255, 255, 255]$ is the spectral radiation intensity for a typical 8bit RGB camera [135].

Only the DAB channel image $I_{_{DAB}}$ from the three colour deconvolution output channels is used, which describes the DAB stain according to the chroma difference.

Most previous works set a single threshold on $I_{_{DAB}}$ to separate positively stained tissues. However, as shown in Figure 6.3, the deeply stained positive nuclei can have dark and light pixel values on the DAB channel image, since the strongly stained pixels will have significantly broader hue spectrum. Furthermore, as illustrated in Figure 6.4, the same DAB channel value can correspond to different pixel colours. Also, from Figure 6.4, it is clear that in order to separate the positive stain (brown colour) from the negative stain (blue colour), the DAB channel thresholds should be set based on the luminance values. In this chapter, we use the Luminance Adaptive Multi-Thresholding (LAMT) method developed by the authors [39] to classify positively stained pixels. I_l is the luminance image of the original RGB image calculated according to Rec. 709:

$$I_l = 0.2126 \times I_R + 0.7152 \times I_G + 0.0722 \times I_B. \tag{6.6}$$

The transformed pixels are thresholded with different values according to its luminance instead of a single threshold, the threshold t_k is assigned as follows:

$$t_k = \underset{c \in \mathscr{C}}{\operatorname{argmax}} P(c | I^k_{DAB}(m,n)) \tag{6.7}$$

where $\mathscr{C} = \{c_{DAB}, c_H\}$ is the stain label.

Luminance Adaptive Random Forest (LARF) is not used for this work. The reason is that LARF requires a number of accurately labelled images for training, while the digital TMA images we utilized in this chapter have no pixel-level positive stain labels.

Once we have separated the positive stain from the negative stain, we need to find a way to describe the stain intensity. As we have already seen in Figure 6.3 and Figure 6.4, the pixel values of I_{DAB} can not describe the biomarker stain intensity. We propose to use a scheme described in Eq.6.8 to assign stain intensity values to pixels:

$$I_{la}(m,n) = \begin{cases} I_l(m,n) & \text{if } I_{DAB}(m,n) \text{ is positive} \\ 255 + (255 - I_l(m,n)) & \text{if } I_{DAB}(m,n) \text{ is negative} \end{cases} \tag{6.8}$$

where $I_{la}(m,n)$ is the stain intensity description image.

The idea is that for the positive stain pixels, $I_{la}(m,n)$ is the same as the luminance component of the original image in order to preserve the morphology of the positive nuclei; for the negative stain pixels, $I_{la}(m,n)$ will have a higher value for strongly stained pixels (darker blue colour) and a lower value for weakly stained pixels (lighter blue colour). In order to separate the positive and negative pixel values clearly, we add an offset of 255 to the negatively stained pixels (most negative stain pixels will have a high $I_l(m,n)$ and positive stain pixels will have a low $I_l(m,n)$, the value of positive and negative pixels will be clearly separated in $I_{la}(m,n)$). Therefore, the larger $I_{la}(m,n)$ is, the weaker is the stain, the smaller $I_{la}(m,n)$ is, the stronger is the stain. When $I_{la}(m,n)$ is below or equal to 255, it is a positive stain pixel. In this way, we have obtained an image which gives a continuous description of the stain intensity of the image. Instead of setting artificial boundaries to separate the different degrees of stain intensity, we have now a continuous description of the stain intensity (see Figure 6.5). Note that the pixel values of the final image are normalized to the range from 0 to 1.

6.2.2 Nuclei and tumour maps

As discussed above, the important information pathologists use to come up with the H-Score is the number of nuclei and the number of tumour nuclei in the TMA image. Therefore, we need to extract these two pieces of information and we use two separate FCNs, one for segmenting all nucleus and the other for segmenting tumour nucleus only.

To segment the tumour region, we use our own manually pixel-wise labelled tumour TMA images to train the FCN. While for segmenting general nuclei which detect both tumour and non-tumour nuclei, we use a transfer learning strategy to train another FCN. For general nuclei detection, the training data is obtained from three different datasets: immunofluorescence (IIF) stained HEp-2 cell dataset [136], Warwick hematoxylin and eosin (H&E) stained colon cancer dataset [84], and our own DAB-H TMA images. Since these three image sets are stained with different types of biomarkers, we transform the colour image into grayscale for training. Training on a mixed image set could help to reduce overfitting on limited medical datasets and further boost the performance and robustness [83].

For both the general nuclei detection network and the tumour nuclei detection network, we use the symmetric U shape network architecture (U-Net) [101] with skip connection. The high resolution features from the contracting path are combined with the output from the upsampling path, which allows the network to learn the high resolution contextual information. Dice coefficient, also known as Dice similarity coefficient (DSC), is widely used as a statistical validation metric to evaluate the performance of image segmentation:

$$DSC = \frac{2|X \bigcap Y|}{|X|+|Y|}, \tag{6.9}$$

where $|X|$ and $|Y|$ are the cardinalities of the two sets. So, the loss function is designed as:

$$L_{mask} = -log \frac{2\sum_{m,n} \omega\tau}{\sum_{m,n} \omega^2 + \sum_{m,n} \tau^2}, \tag{6.10}$$

where ω is the predicted pixel and τ is the ground truth.

6.2.3 The H-Score prediction framework

The overview of the H-Score prediction framework is illustrated in Figure 6.6. It consists of three stages: 1) Nuclei segmentation, tumour segmentation, and stain intensity description; 2) Constructing the Stain Intensity Nuclei Image (SINI) and the Stain Intensity Tumour Image (SITI); 3) Predicting the final histochemical score (H-Score) by the Region Attention Multi-column Convolutional Neural Network (RAM-CNN). The rationale of this architecture is as follows: as only the number of nuclei, the number of tumour nuclei and the stain intensity of the tumour nuclei are the useful information for predicting H-Score, we therefore first extract this information. Rather than setting artificial boundaries for

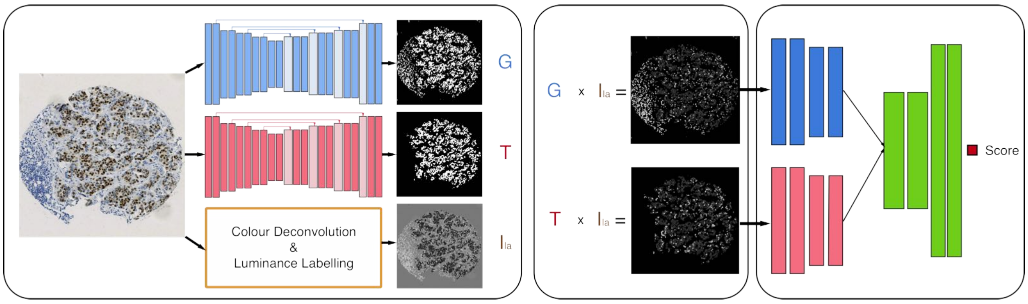


Figure 6.6: The overview of our proposed H-Score prediction framework. An input TMA image is first processed by two FCNs to extract tumour cells and all cells (tumour and non-tumour) to produce two mask images. The input image is also processed by colour deconvolution and positive stain classification to output a stain intensity description image. The two mask images are used to filter out irrelevant information in the stain intensity description image and only the useful information is fed to a deep convolutional neural network for the prediction of H-Score of the input TMA.

the categories of stain intensity, we retain a continuous description of the stain intensity. Only the information useful for predicting the H-Score is presented to a deep CNN to estimate the H-Score of the input image. This is in contrast to many methods in the literature where the whole image is thrown to the CNN regardless if a region is useful or not for the purpose.

The detail of the first stage have been described in Section 6.2.1 and 6.2.2. As illustrated in Figure 6.6, an input TMA image $I(m,n)$ is processed by the tumour detection network which will output a binary image mask, $T(m,n)$, marking all the tumour nuclei, where $T(m,n) = 1$ if $I(m,n)$ is a part of a tumour nuclei and $T(m,n) = 0$ otherwise; by the general nuclei detection network which will output another binary image mask, $G(m,n)$, marking all tumour and non-tumour nuclei, where $G(m,n) = 1$ if $I(m,n)$ is a part of a nuclei and $G(m,n) = 0$ otherwise; and by the colour deconvolution and stain intensity labelling operation of Equation (8) to produce the stain intensity description image $I_{la}(m,n)$. In the second stage, we construct SINI and SITI by multiplying the nuclei mask image $G(m,n)$ and tumour mask image $T(m,n)$ with the stain intensity description image $I_{la}(m,n)$, i.e. $SINI = I_{la}(m,n) \times G(m,n)$, and $SITI = I_{la}(m,n) \times T(m,n)$. Hence, all background pixels are zero, while only the region of interests (ROI) are retained in SINI and SITI. All necessary information is preserved for histochemical assessment. Removing the background and only retaining ROI will enable the RAM-CNN convolutional layers to focus on foreground objects [137] which will significantly reduce computational costs and improve performance.

<table>
<tr><th>Layer</th><th colspan="2">Input / Filter Dimensions</th></tr>
<tr><td>Input</td><td>$512 \times 512 \times 1$</td><td>$512 \times 512 \times 1$</td></tr>
<tr><td>Conv1
MaxPooling</td><td>$8 \times 7 \times 7$
2×2</td><td>$8 \times 7 \times 7$
2×2</td></tr>
<tr><td>Conv2
MaxPooling</td><td>$16 \times 5 \times 5$
2×2</td><td>$16 \times 5 \times 5$
2×2</td></tr>
<tr><td>Conv3
MaxPooling</td><td colspan="2">$64 \times 3 \times 3$
2×2</td></tr>
<tr><td>Conv4
MaxPooling</td><td colspan="2">$64 \times 3 \times 3$
2×2</td></tr>
<tr><td>FC1</td><td colspan="2">2048</td></tr>
<tr><td>FC2</td><td colspan="2">1024</td></tr>
</table>

Table 6.1: The architecture of Region Attention Multi-channel Convolutional Neural Network (RAM-CNN).

The proposed RAM-CNN is a deep regression model with dual input channels. The architecture of RAM-CNN is shown in Table 6.1. Two inputs correspond to SINI and SITI respectively, and the input size is 512×512. The parameters of the two individual branches are updated independently for extracting cell and tumour features respectively, without interfering with each other. The two pipelines are merged into one after two convolutional layers for H-Score prediction. The loss function for H-Score prediction is defined as:

$$L_{score} = \frac{1}{N} \sum_{i=1}^{N} \|F_{RAM}(SINI_i, SITI_i) - l_i\|_2, \tag{6.11}$$

where $F_{RAM}(SINI_i, SITI_i)$ is the estimated score generated by RAM-CNN. l_i is the ground truth H-Score.

6.3 Experiments and Results

6.3.1 Dataset

The H-Score dataset used in our experiment contains 105 TMA images of breast adenocarcinomas from the NPI+ set [110]. Each image contains one whole TMA core. The tissues are cropped from a sample of one patient which are stained with three different nuclei activity biomarkers: ER, p53, and PgR. The original images are captured at a high resolution of 40× optical magnification, and then resized to 1024×1024 pixels. The dataset is manually marked by two experienced pathologists with H-Score based on common practice. For each TMA core, the pathologists give the percentage of nuclei of different stain intensity levels, and then calculate the H-Score using Eq.6.1. The final label (H-Score) is determined by averaging two pathologists' scores, if the difference between two pathologists is smaller than 20. The dataset is available from the authors on request.

For training the general nuclei detection network, we transform Warwick H&E colon adenocarcinoma [84] and NPI+ images to grayscale; the green channel was extracted from HEp-2 cell dataset [136]. As HEp-2 cell images are IIF stained, the gray value should be inversed.

6.3.2 Data and label augmentation

As in typical medical imaging applications, the dataset sizes are relatively small. In developing deep learning based solutions, it is a common practice to augment the training dataset for training. The training images for general nuclei detection network and tumour detection network are augmented by randomly cropping sub-images as input samples. For the H-Score dataset, rotation with random angles and randomly shifting the image horizontally and vertically within 5% of image height and width are performed to augment the training set.

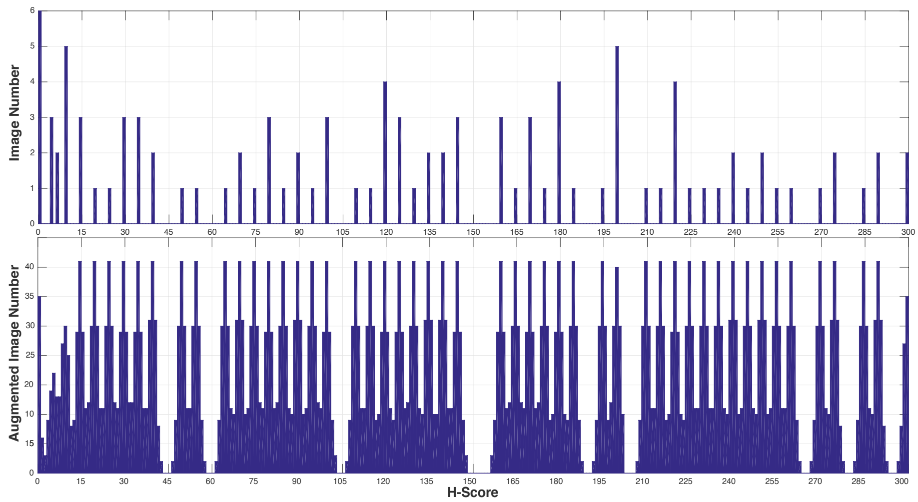


Figure 6.7: The top graph is the original dataset label histogram; The bottom is the augmented label histogram.

As shown in the top row of Figure 6.7, the distribution of the label (H-Score) in the original dataset is unbalanced, some labels (H-Scores) have far more samples than others. Furthermore, one of the biggest problems is that because we have only limited number of samples, the H-Score values are discrete and discontinuous. There are many gaps between two H-Scores that have no data. Also, the values of the TMA image score given by the pathologists have a quantitative step-size of 5. Therefore, if an image has a score of 125, it means it has a value of around 125, the values in the vicinity of 125, i.e., 126 or 124 should also be suitable for labelling that image. In order to solve the ambiguity issue, we introduce Distributed Label Augmentation (DLA) which was inspired by the work of [138, 139].

In the traditional regression method, a given dataset $\{(I_1,l_1),(I_2,l_2),\cdots,(I_D,l_D)\}$ pairs the instance I_d for $1 \le d \le D$ with one single l_d from the finite class label space $\mathscr{L} = \{l^0,l^1,\cdots,l^C\}$, where C is the label size (*e.g.*, $C = 301$ for H-Score). In this paper, the label is augmented so that one instance is associated with a number of labels. Formally, the dataset can be described as $\{(I_1,Y_1),(I_2,Y_2),\cdots,(I_D,Y_D)\}$, and $Y_d \subseteq \mathscr{Y}$ is a set of labels $\{y_d^{(1)},y_d^{(2)},\cdots,y_d^{(S)}\}$, where S is the augmented label number for I_d. $y_d^{(s)}$ is sampled repeatedly from $\mathscr{L}$ based on a probability density function of following Gaussian distribution:

$$p(y_d^{(s)} = l^c) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(l^c - \mu)^2}{2\sigma^2}) \tag{6.12}$$

where μ is the mean which equal to l_d and σ is standard deviation. Thus, $\sum_{s=1}^{S} p(y_d^{(s)}) = 1$ for each original TMA image. Consequentially, for an image x_i from the augmented training set, its ground truth labels are assigned by repeatedly sampling from $\mathscr{L}$ according to Eq.6.12. The augmented label histogram is shown at the bottom row of Figure 6.7.

6.3.3 Implementation details

The network architecture for both the tumour nuclei detection and general nuclei detection models is the same as the U-Net [101] with an input size of 224×224. The filter size of tumour detection net is half narrower than that of general cell detection net. All networks use rectified linear unit (ReLU) activation function for the convolutional layer. The final cell and tumour region maps are predicted using the sliding window.

A leave 5 out cross validation strategy is used for RAM-CNN model training, which means that in each round of testing, we randomly sample 5 TMAs as testing and the other 100 TMAs as training images. As explained previously, the training set is augmented via rotation and shift. The images are firstly resized to 512×512 before fed into the RAM-CNN. We set $\sigma = 0.9$ to generate the H-Score distribution for ground truth label augmentation. We also add dropout layers after two fully connected layers with the rates of 0.3 and 0.5 respectively. The regression network is optimized by Adam [140] with an initial learning rate of 0.001.

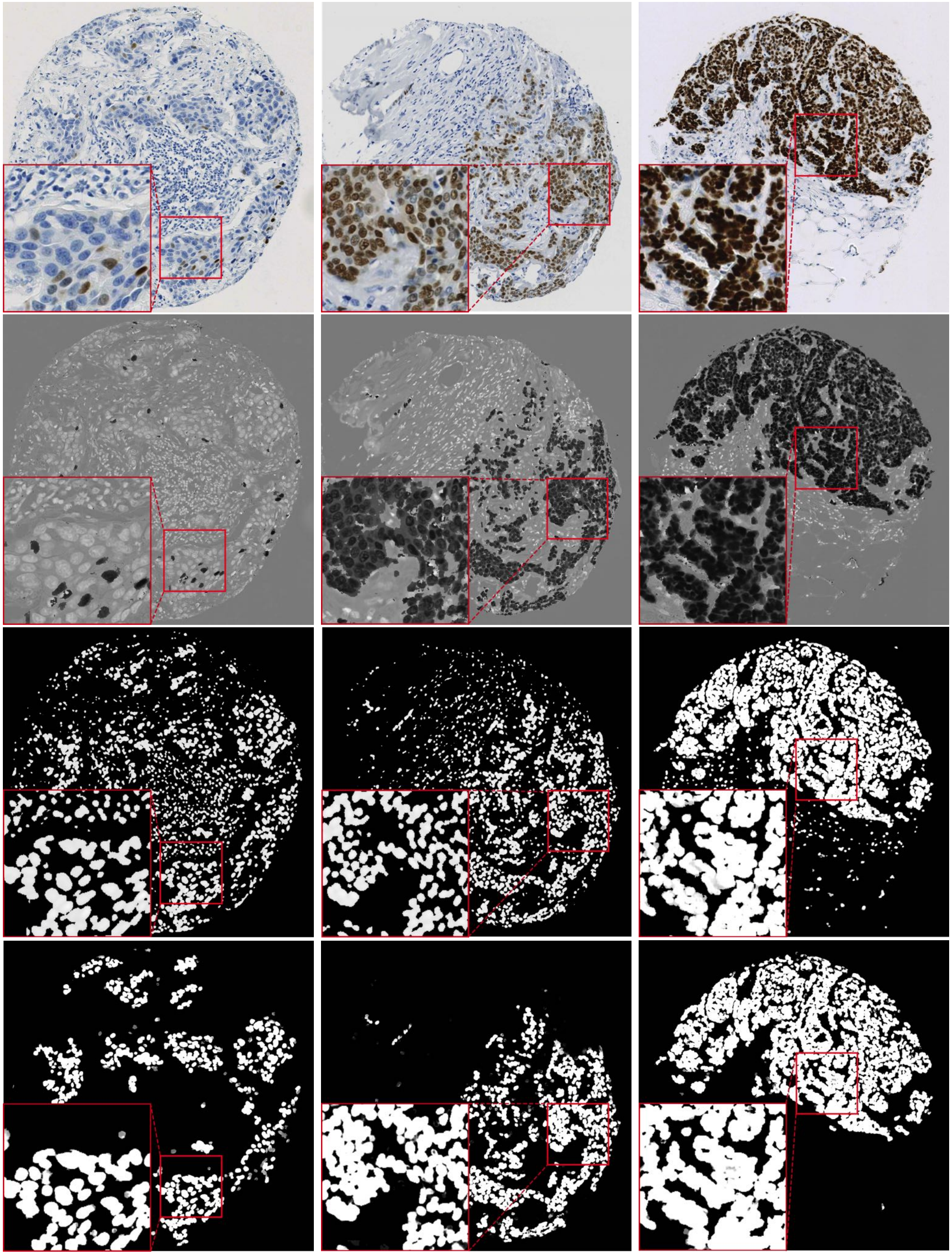
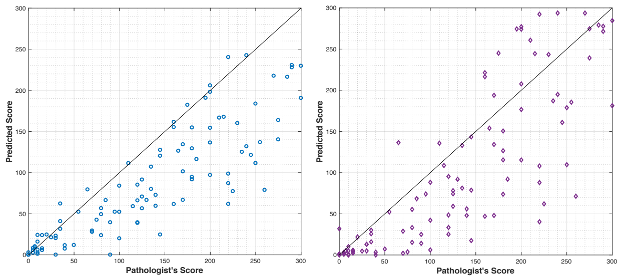


Figure 6.8: Examples of intermediate images in the automatics H-Score prediction pipeline. From top to bottom: the original RGB image, luminance labelled stain intensity image, nuclei mask image, and tumour mask image respectively.

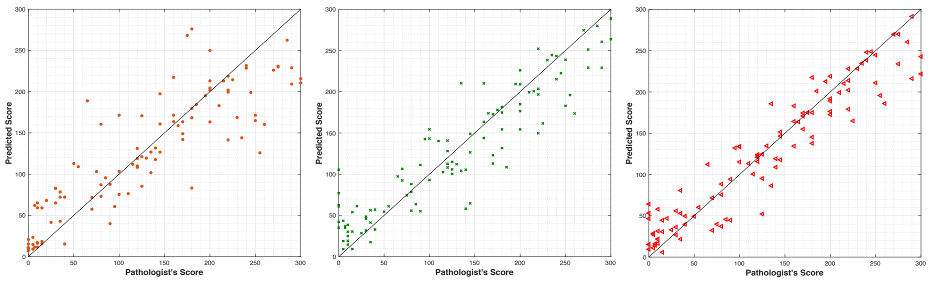
6.3.4 Results and discussions

Experimental Results

Figure 6.8 shows some examples of the intermediate images in the automatic H-Score prediction pipeline. It is seen that the luminance labelled stain intensity image marks a sharp distinction between positive and negative stains. This shows that our maximum a posteriori (MAP) classifier based Luminance Adaptive Multi-Thresholding (LAMT) method [39] can reliably separate positive DAB stains for a variety of images. It also shows that our stain intensity labelling strategy can preserve the morphology of the nuclei, separate positive and negative stains while retaining a continuous description of the positive stain intensity.



(a) NAP (b) NNP



(c) RGB-CNN (d) RA-CNN (e) RAM-CNN

Figure 6.9: Scatter plots of the predicted scores of different models vs the pathologists' manual scores.

Figure 6.10 shows the training curves of the Dice coefficient for the general nuclei detection network and the tumour nuclei detection network respectively.

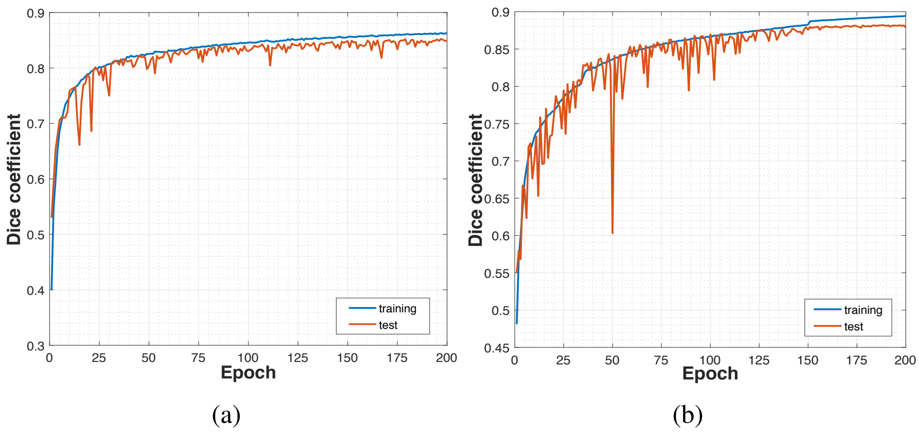


Figure 6.10: Training results for general nuclei detection network (a) and tumour nuclei detection network (b).

Both networks converged after 170 epochs respectively. The nuclei mask images (see Figure 6.8) show that the deep convolutional network trained with mixed datasets using transfer learning can successfully detect the nuclei in our H-Score dataset. The tumour segmentation network is able to identify tumour region from normal tissues. It is worth noting that the ground truth masks for the two detection networks are different. All nucleus in Warwick colon cancer images are labelled with circular masks with a uniform size, while tumour region masks are pixel level labelled. Therefore, the final predicted maps generated by the two networks for the same nucleus are different. In addition, it is found that the mask dilation becomes evident with the increase of DAB stain intensity. One possible reason is that the strong homogeneous stain makes the nuclei texture and edge feature difficult to extract.

To evaluate the performance of our proposed RAM-CNN and the two H-Score relevant images SINI and SITI, we compare our model with two traditional single input pipeline CNNs: RGB-CNN and RA-CNN (i.e., region attention CNN). The RGB-CNN takes the original RGB TMA image with the shape of $512 \times 512 \times 3$ as input, and output the H-Score prediction. To investigate the effect of multi-column architecture, we combine SINI and SITI as a two channel image of $512 \times$

512×2 for the input of RA-CNN. The architectures of RGB-CNN and RA-CNN are the same as a single pipeline RAM-CNN (see Table 6.1).

We also calculate the H-Score using Eq.6.1 based on the nuclei area percentage (NAP) and nuclei number percentage (NNP). Specifically, the luminance labelled stain intensity description image I_{la} is first calculated according to the description in Section 6.2.1. The pre-defined thresholds [4] are utilized for categorizing the pixels into different DAB-H stain intensity groups. For the NAP method, the predicted H-Score is calculated according to the percentages of the area in different stain intensity groups. NNP employs the NIH ImageJ tool [141] with a multi-stage cell segmentation technique [142] for cell detection. The detected cells are classified into *unstained*, *weak*, *moderate*, and *strong* groups using the pre-defined thresholds for H-Score calculation.

Model	MAE	SD	CC	P value
NAP	47.09	46.03	0.87	< 0.001
NNP	46.48	55.18	0.82	< 0.001
RGB-CNN	32.01	44.46	0.87	< 0.001
RA-CNN	27.22	35.72	0.92	< 0.001
RAM-CNN	**21.33**	**29.14**	**0.95**	< 0.001
Human	20.63	30.55	0.95	< 0.001

Table 6.2: Performance comparison with different regression models. The last line (Human) are difference between the H-Scores given by the two pathologists.

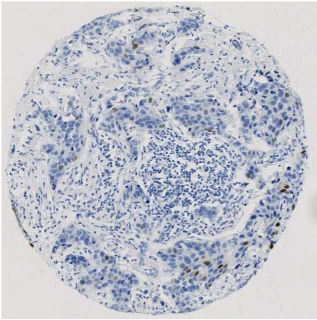
In this chapter, Mean Absolute Error (MAE), Standard Deviation (SD) and the correlation coefficient (CC) between the predicted H-Score and the average H-Score of the two pathologists are used as the evaluation metrics. As a reference, we also calculate the MAE, SD and CC between the H-Scores given by the two pathologists of all original diagnosis data. Results are shown in Table 6.2.

As can be seen, the NAP based prediction gives the highest MAE with large deviations in the cross validation, which is followed by NNP. Our RAM-CNN framework achieves the lowest prediction error (21.33); a traditional CNN setting with the proposed SINI and SITI as input gives the second lowest prediction error

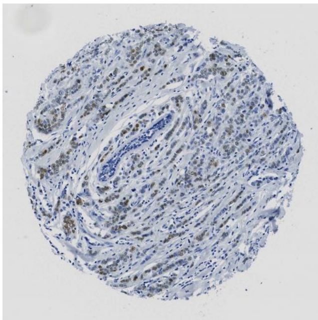
(27.22). This verifies the effectiveness of our proposed approach to filtering out irrelevant pixels and only retain H-Score relevant information in SINI and SITI. All deep learning based methods outperform NAP and NNP by a large margin. To investigate the statistical significance of automatically predicted H-Scores, the correlation of the predicted and those of the pathologists scores and its P value are also calculated. The correlation between pathologists' scores and those predicted by RAM-CNN is 0.95 with a P value of < 0.001 which means there is strong evidence against the null hypothesis [143].

It is interesting to observe that the difference between our RAM-CNN predicted H-Scores and the average of the two pathologists H-Scores (MAE = 21.33, SD = 29.14, CC = 0.95) are on par with the difference between the two pathologists (MAE = 20.63, SD = 30.55, CC = 0.95). While the MAE between the RAM-CNN and humans is slightly higher than that between humans, the SD between humans is higher than that between RAM-CNN and humans. The CC between humans and machine and that between humans are the same.

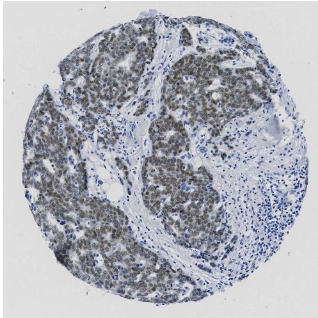
Figure 6.9 illustrates the scatter plots between the model predicted scores and the pathologists' scores. Most of the predicted scores of NAP are lower than the ground truth. At the lower end, NNP predicted scores are lower than the ground truth while at the higher end the predicted scores are higher than the ground truth. These two methods are affected by several low-level processing components including the pre-defined stain intensity thresholds and the nuclei segmentation accuracy. Our proposed framework gives more accurate prediction results compared to traditional single pipeline CNN, further demonstrating that imitating the pathologists' H-Scoring process by only keeping useful information is an effective approach.



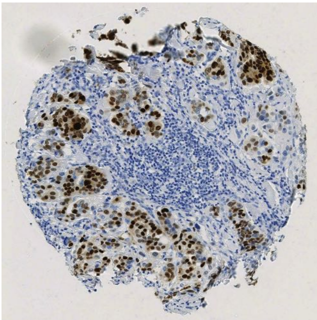
(a) 0-49



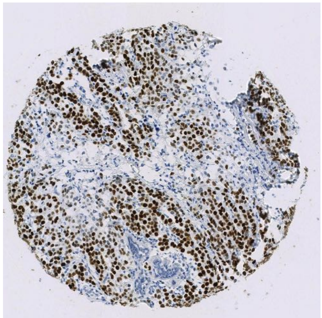
(b) 50-99



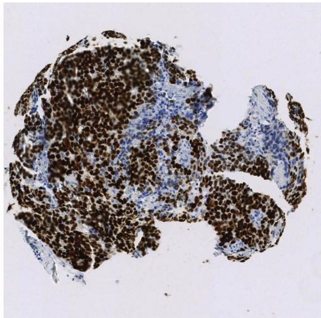
(c) 100-149



(d) 150-199



(e) 200-249



(f) 250-300

Figure 6.11: Example TMA images extracted from different H-Score groups.

Discussions

In this paper, we introduced an end-to-end system to predicted H-Score. To investigate the reason for scoring discrepancy between the proposed algorithm and the pathologists, we firstly compare the H-Score prediction results for different biomarkers as shown in Table 6.3. The proposed framework gives the best accuracy in all three biomarker images. The performances are slightly different for different biomarkers. This is to be expected because different markers will stain the tissues differently. Although the difference is not large, whether it will be useful to train a separate network for different biomarkers is something worth investigating in the future.

To see how the algorithms perform differently across the dataset, we divide the TMA images into 6 groups according to their pathologists' scores. Example

Biomarker	ER	P53	PgR
No. of TMA	32	33	40
NAP	42.02	50.68	48.17
NNP	43.53	46.72	48.92
RGB-CNN	24.90	31.57	38.19
RA-CNN	25.43	23.39	31.82
RAM-CNN	21.01	16.66	25.44

Table 6.3: Comparing MAE of different methods on three different biomarkers.

TMA images of each group are illustrated in Figure 6.11. For each group, we count the number of TMAs with absolute error (AE) smaller than 10, between 10 and 30, and larger than 30 respectively. The results of different methods are shown in Figure 6.12. It is seen that in the low H-Score group of 0-49, traditional methods of NAP and NNP give more accurate predicted scores than CNN based methods. It is found that most low score TMAs are unstained or weakly stained as shown in Figure 6.11(a). The accurate predictions from NAP and NNP indicate that the predefined threshold for separating *unstained* and *weak* (see Figure 6.5) is compatible with pathologists' criteria. The deep learning based methods do not set stain intensity thresholds explicitly and their performances across the six groups are relatively even.

The accuracies of NAP and NNP decrease rapidly with the increase of the H-Score. As shown in Figure 6.11, the stain intensity and image complexity increase with the H-Score which directly affect the performance of traditional methods. The result also indicates that the pre-defined stain intensity thresholds for *moderate* and *strong* classes (see Figure 6.5) are less compatible with the pathologists' criteria. Furthermore, the large coefficients of *moderate* and *strong* stain (see Eq.6.1) would magnify the errors of area and nuclei segmentation in NAP and NNP respectively.

Three deep learning based methods give worse results on the groups with fewer images (i.e., group 50-99 and 250-300), which indicates the importance of a

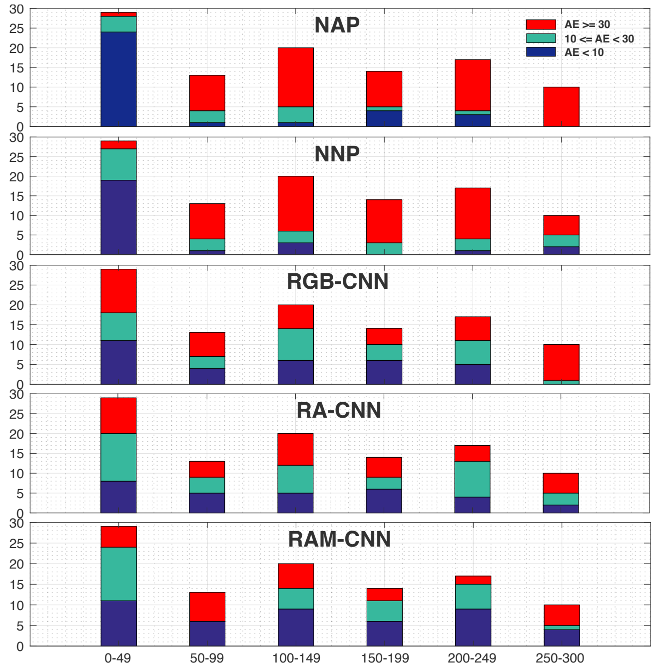
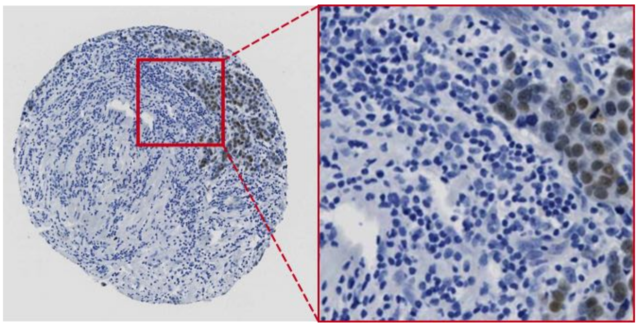


Figure 6.12: Comparison of performances of different methods in different H-Score groups.

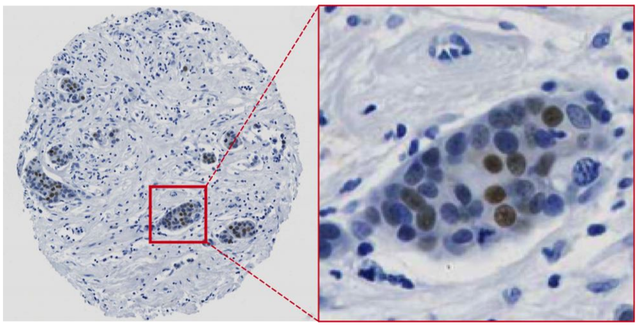
large training data size. In addition, the uneven distribution of original dataset may also affect the predicted accuracy.

We further analyse the TMAs individually to investigate the effect of image quality on the proposed algorithm. We found that for those TMAs where the tissues are clearly stained, and the cellular structure is clear without severe overlap (see Figure 6.13), our algorithm can give a very accurate prediction. On the other hand, poor image quality causes errors. In the images that are most easily mis-scored by our algorithms, we found three significant characteristics as shown in Figure 6.14.

The TMA core in Figure 6.14(a) contains large out-of-focus regions, which happens more commonly on strongly stained tissues. The blur regions directly affect the performance of nuclei segmentation, as well as the nuclei and tumour detection accuracy. They also hinder the final regression network from extracting topological and morphological information.



(a)



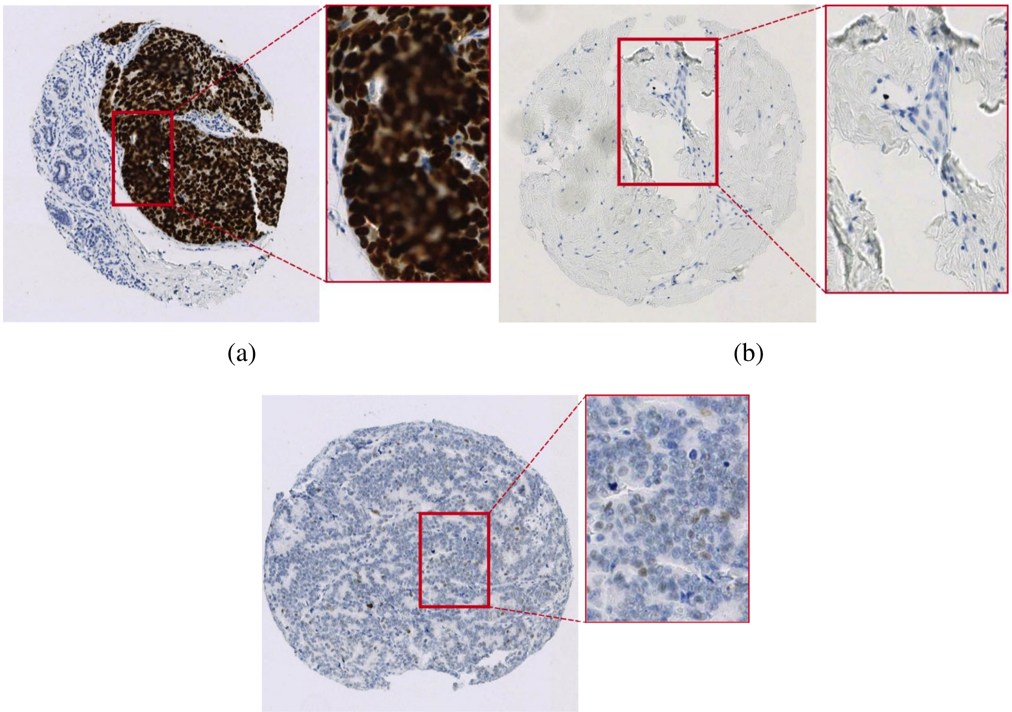
(b)

Figure 6.13: Examples of accurately scored TMAs by proposed algorithm. The absolute errors generated by RAM-CNN of both (a) and (b) are smaller than 2.

Tissue folds (see Figure 6.14(b)) occurs when a thin tissue slice folds on itself, and it can happen easily during slide preparation especially in TMA slides. Tissue-fold would cause out-of-focus during slide scanning. Furthermore, a tissue fold in a lightly stained image can be similar in appearance to a tumour region in a darkly stained image [144]. Hence, the segmentation accuracy of colour deconvolution would be greatly affected in tissue-fold regions.

Heterogeneity and overlapping as shown in Figure 6.14(c) also affect the automatic scoring performance. The stain heterogeneity gives rise to a large discrepancy of stain intensity in a single nucleus, and nuclei overlapping adds to the difficulty.

These three difficulties directly affect the predicted results of the proposed method, and we found that most large mis-scored TMAs contain one or more of these



(a) (b) (c)

Figure 6.14: Examples of sources of big scoring discrepancy between algorithm and pathologist. (a) out of focus; (b) tissue folds; (c) heterogeneity and overlapping.

characteristics. We found that there were 9 low image quality TMAs in our dataset and if we exclude these 9 lowest-quality TMA images, the average MAE of our RAM-CNN is 18.79. Therefore, future works need to overcome these issues in order to achieve a high prediction performance. To solve the problem of out-of-focus, heterogeneity and overlapping, adding corresponding images in the training set to promote robustness is one potential quality assurance methods. In addition, the deep learning based scoring system can be developed to add nuclei number estimation function for accurate assessment. It is also necessary to add the function of automated detection and elimination of tissue-fold regions before H-Score assessment.

6.4 Summary

In this chapter, we have developed a deep learning framework for automatic end-to-end H-Score assessment for breast cancer TMAs. Experimental results show that automatic assessment for TMA H-Score is feasible. The H-Scores predicted by our model have a high correlation with H-Scores given by experienced pathologists. We show that the discrepancies between our deep learning model and the pathologits are on par with those between the pathologists. We have identified image out-of-focus, tissue fold and overlapping nuclei as the three major sources of error. We also found that the major discrepancies between pathologists and machine predictions occurred in images that will have a high H-Score value. These findings have suggested future research directions for improving accuracy.

Chapter 7

Concluding Remarks

In this thesis, problems related to digital histopathology images have been studied. Its marjor contribution will now be summarized, followed by a discussion of some possible research directions related to these studies for future work.

7.1 Main Contributions

The first contribution in Chapter 3 is to propose the luminance adaptive multiple threshold (LAMT) method for DAB stain separation. The mistake in most existing stain separation approaches, in which only a chroma difference is used, without taking into consideration the effect of luminance on DAB stain was pointed out. The experimental results shows that the proposed method in the present thesis is able to dramatically promote the accuracy of this separation, and that the proposed combination of LAMT combined with random forests outperforms all existing methods.

The second contribution is to the cell classification task: two different models have been proposed. In the first, which is a model based on hand-crafted features, as is traditional in the subject, an objectgraph feature was introduced to describe patterns in the morphology of HEp-2 cells. Unlike normal approaches based on hand-crafted feature based, the proposed model transforms the foreground region to different sizes of circles, designed according to the HEp-2 cell pattern. The experiments show that the circle-transform feature could achieve promising results.

The second proposal was a model for cell classification based on a convolutional neural network. This proposed model uses a multi-task learning strategy with alternative training to predict the cell pattern class and reconstruct the original image. A series of experiments has been conducted, and showed that the proposed model is able to significantly improve the classification performance when the training dataset is small, and mitigate the effect of overfitting. Further, this proposed model outperforms all state-of-the-art methods on the two HEp-2 I3A datasets and CRC dataset, which means the cell classification CNN model does not need a very deep and complex model, but can maximize the use of the training images by extracting more information for the classification.

Chapter 6 introduces a novel end-to-end framework for automatic H-Score scoring. As an image-level assessment, it is, as far as the present author's knowledge extends, the first end-to-end system that takes TMA images and directly predicts a clinical score. To extract the DAB stain intensity information, it uses a luminance image assigned after an adaptive luminance thresholding. Since the dataset is labelled ambiguously with quantitative steps, there was introduced a distributed label augmentation, which labelled single image with a set of scores following a Gaussian distribution. Experimental results show that automatic assessment of a TMA H-Score is feasible and is highly correlated with pathologists' scores.

7.2 Limitations and Future Research

The research presented in this thesis can be considered as just a beginning in each of the respective fields. They could not yet perfectly solve the corresponding problems, and there is room for future improvement.

For stain separation, recent work, including the method proposed here, still require pixel-level labels. However, pixel-level labelling is time consuming, and the accuracy of such labels directly affects the final separation performance. Hence, future work should focus on weak supervised learning or unsupervised learning to reduce the effect of human labelling. In addition, while studying this problem, it was found that there is no standard quantitative criterion for describing the DAB stain intensity.

For cell or nucleus classification, future research should focus on deep learning techniques, because of their promising performance. As various new CNN models have been proposed for natural image classification, more models could be introduced to the field of digital histpathology. Furthermore, increasing the speed of the classification model would also be a good direction.

The proposed end-to-end deep learning based model is the first high-level CAD system for TMA H-Score assessment. Therefore, there are still some problems, such as big scoring discrepancies between the algorithm and pathologists in TMA images with out-of-focus, tissue-fold, and serious overlapping. Further, the dataset used here only contains 105 TMA cores. In the future, more data should be collected to create a new big dataset to evaluate the new system. To improve the performance of the final score prediction, a more advanced model is required to increase the general accuracy of the detection of the nuclei and tumour nuclei.

7.3 Summary

In summary, the present thesis has investigated three typical problems in digital histopathology. The problems can be categorized into three levels, according to a histopathological perspective: pixel level, cell level, and image level. The pixel-level DAB stain separation method introduced in Chapter 3 belongs to the study of the colour of pathology stains. The stain colour is the most basic, and one of the most important pieces of information for diagnosis. At the cell level, one aspect of cell classification has been studied here, and two different models have been proposed. Future work would try to investigate more cell-level tasks such as cell segmentation and cell counting, and try to solve them simultaneously. Finally, at the image level, this proposed framework is able to combine both colour and morphology information for a final comprehensive assessment. This research has demonstrated the possibility of using deep learning techniques to automatically predict the clinical scores.

From the machine learning and computer vision perspective, the DAB stain separation model is a two-class classification method, according to the 3-dimensional dataset of RGB values, whereas cell image classification belongs to the multiple class classification category. The latter method can be separated into two parts: object detection for nuclei region segmentation, and regression for H-Score prediction. While the research for this thesis was been carried out, there was occurring the revolution of deep learning. Hence, the latter two methods use deep convolutional neural network models. Future work will try more advanced models and techniques to develop a digital histopathology image analysis system.

Bibliography

[1] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution.," *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, vol. 23, no. 4, pp. 291–299, 2001.

[2] E. M. Brey, Z. Lalani, C. Johnston, M. Wong, L. V. McIntire, P. J. Duke, and C. W. Patrick, "Automated selection of dab-labeled tissue for immunohistochemical quantification," *Journal of Histochemistry & Cytochemistry*, vol. 51, no. 5, pp. 575–584, 2003.

[3] R. Fu, X. Ma, Z. Bian, and J. Ma, "Digital separation of diaminobenzidine-stained tissues via an automatic color-filtering for immunohistochemical quantification," *Biomedical optics express*, vol. 6, no. 2, pp. 544–558, 2015.

[4] N. Trahearn, Y. W. Tsang, I. A. Cree, D. Snead, D. Epstein, and N. Rajpoot, "Simultaneous automatic scoring and co-registration of hormone receptors in tumor areas in whole slide images of breast cancer tissue slides," *Cytometry Part A*, 2016.

[5] L. C. Junqueira and J. Carneiro, *Basic histology text and atlas*. London: McGraw Hill, 2005, 2005.

[6] L. He, L. R. Long, S. Antani, and G. Thoma, "Computer assisted diag-

nosis in histopathology," *Sequence and genome analysis: methods and applications*, vol. 510, pp. 271–287, 2010.

[7] 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.

[8] E. Gardner, "Image analysis in histology. conventional and confocal microscopy," *Journal of Microscopy*, vol. 182, no. 3, pp. 235–236, 1996.

[9] J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature medicine*, vol. 4, no. 7, pp. 844–847, 1998.

[10] D. Pilla, F. M. Bosisio, R. Marotta, S. Faggi, P. Forlani, M. Falavigna, I. Biunno, E. Martella, P. De Blasio, S. Borghesi, *et al.*, "Tissue microarray design and construction for scientific, industrial and diagnostic use," *Journal of pathology informatics*, vol. 3, 2012.

[11] A. Nocito, J. Kononen, O.-P. Kallioniemi, and G. Sauter, "Tissue microarrays (tmas) for high-throughput molecular pathology research," *International journal of cancer*, vol. 94, no. 1, pp. 1–5, 2001.

[12] J. J. Sciubba, U. C. O. S. Group, *et al.*, "Improving detection of precancerous and cancerous oral lesions: computer-assisted analysis of the oral brush biopsy," *The Journal of the American Dental Association*, vol. 130, no. 10, pp. 1445–1457, 1999.

[13] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, "Hematoxylin and eosin staining of tissue and cell sections," *Cold Spring Harbor Protocols*, vol. 2008, no. 5, pp. pdb–prot4986, 2008.

[14] K. R. Katikireddy and F. OSullivan, “Immunohistochemical and immunofluorescence procedures for protein analysis,” *Gene Expression Profiling: Methods and Protocols*, pp. 155–167, 2011.

[15] J. Duraiyan, R. Govindarajan, K. Kaliyappan, and M. Palanisamy, “Applications of immunohistochemistry,” *Journal of pharmacy & bioallied sciences*, vol. 4, no. Suppl 2, p. S307, 2012.

[16] G. G. Kovacs, *Neuropathology of Neurodegenerative Diseases Book and Online*. Cambridge University Press, 2014.

[17] E. C. Jensen, “Quantitative analysis of histological staining and fluorescence using imagej,” *The Anatomical Record*, vol. 296, no. 3, pp. 378–381, 2013.

[18] N. Farahani, A. V. Parwani, and L. Pantanowitz, “Whole slide imaging in pathology: advantages, limitations, and emerging perspectives,” *Pathol Lab Med Int*, vol. 7, pp. 23–33, 2015.

[19] 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.

[20] R. Gutiérrez, F. Gómez, L. Roa-Peña, and E. Romero, “A supervised visual model for finding regions of interest in basal cell carcinoma images,” *Diagnostic pathology*, vol. 6, no. 1, p. 26, 2011.

[21] V. Roullier, O. Lézoray, V.-T. Ta, and A. Elmoataz, “Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization,” *Computerized Medical Imaging and Graphics*, vol. 35, no. 7, pp. 603–615, 2011.

[22] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, “Methods for nuclei detection, segmentation, and classification in digital histopathology:

a reviewcurrent status and future potential," *IEEE reviews in biomedical engineering*, vol. 7, pp. 97–114, 2014.

[23] H.-A. Lehr, T. W. Jacobs, H. Yaziji, S. J. Schnitt, and A. M. Gown, "Quantitative evaluation of her-2/neu status in breast cancer by fluorescence in situ hybridization and by immunohistochemistry with image analysis," *American journal of clinical pathology*, vol. 115, no. 6, pp. 814–822, 2001.

[24] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunyé, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images: A comparative study," *Journal of digital imaging*, vol. 29, no. 4, pp. 496–506, 2016.

[25] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, P. Khurd, A. Kamen, L. Grady, E. Krupinski, A. Graham, *et al.*, "Automated detection of diagnostically relevant regions in h&e stained digital pathology slides," in *Proc. SPIE*, vol. 8315, 2012.

[26] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, 2013.

[27] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[28] H. Masmoudi, S. M. Hewitt, N. Petrick, K. J. Myers, and M. A. Gavrielides, "Automated quantitative assessment of her-2/neu immunohistochemical expression in breast cancer," *IEEE transactions on medical imaging*, vol. 28, no. 6, pp. 916–925, 2009.

[29] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, "Big data analytics in bioinformatics: A machine learning perspective," *arXiv preprint arXiv:1506.05101*, 2015.

[30] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artificial Intelligence in Medicine*, vol. 75, pp. 51–63, 2017.

[31] K. McCarty Jr, L. Miller, E. Cox, J. Konrath, and K. McCarty Sr, "Estrogen receptor analyses. correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies.," *Archives of pathology & laboratory medicine*, vol. 109, no. 8, pp. 716–721, 1985.

[32] R. Prati, S. K. Apple, J. He, J. A. Gornbein, and H. R. Chang, "Histopathologic characteristics predicting her-2/neu amplification in breast cancer," *The breast journal*, vol. 11, no. 6, pp. 433–439, 2005.

[33] N. Fedchenko and J. Reifenrath, "Different approaches for interpretation and reporting of immunohistochemistry analysis results in the bone tissue–a review," *Diagnostic pathology*, vol. 9, no. 1, p. 221, 2014.

[34] M. T. Galgano, G. M. Hampton, and H. F. Frierson, "Comprehensive analysis of he4 expression in normal and malignant human tissues," *Modern Pathology*, vol. 19, no. 6, pp. 847–853, 2006.

[35] N. Trahearn, Y. W. Tsang, I. A. Cree, D. Snead, D. Epstein, and N. Rajpoot, "Simultaneous automatic scoring and co-registration of hormone receptors in tumor areas in whole slide images of breast cancer tissue slides," *Cytometry Part A*, vol. 91, no. 6, pp. 585–594, 2017.

[36] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason

grading of histological images," *IEEE transactions on medical imaging*, vol. 26, no. 10, pp. 1366–1378, 2007.

[37] C. R. Taylor, "Immunohistochemistry in surgical pathology: principles and practice," *Histopathology: Methods and Protocols*, pp. 81–109, 2014.

[38] F. Varghese, A. B. Bukhari, R. Malhotra, and A. De, "Ihc profiler: an open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples," *PloS one*, vol. 9, no. 5, p. e96801, 2014.

[39] J. Liu, G. Qiu, and L. Shen, "Luminance adaptive biomarker detection in digital pathology images," *Procedia Computer Science*, vol. 90, pp. 113–118, 2016.

[40] W. Ma and S. Lozanoff, "A full color system for quantitative assessment of histochemical and immunohistochemical staining patterns," *Biotechnic & histochemistry*, vol. 74, no. 1, pp. 1–9, 1999.

[41] E. Kaczmarek, A. Gorna, and P. Majewski, "Techniques of image analysis for quantitative immunohistochemistry," *Rocz Akad Med Bialymst*, vol. 49, no. Suppl 1, pp. 155–158, 2004.

[42] M. Goto, Y. Nagatomo, K. Hasui, H. Yamanaka, S. Murashima, and E. Sato, "Chromaticity analysis of immunostained tumor specimens," *Pathology-Research and Practice*, vol. 188, no. 4, pp. 433–437, 1992.

[43] P. Kohlberger, F. Breitenecker, A. Kaider, A. Lösch, G. Gitsch, G. Breitenecker, and D. Kieback, "Modified true-color computer-assisted image analysis versus subjective scoring of estrogen receptor expression in breast cancer: a comparison.," *Anticancer research*, vol. 19, no. 3B, pp. 2189–2193, 1998.

[44] N.-A. Pham, A. Morrison, J. Schwock, S. Aviel-Ronen, V. Iakovlev, M.-S. Tsao, J. Ho, and D. W. Hedley, "Quantitative image analysis of immunohistochemical stains using a cmyk color model," *Diagnostic pathology*, vol. 2, no. 1, p. 8, 2007.

[45] J. Dong, J. Li, A. Fu, and H. Lv, "Automatic segmentation for ovarian cancer immunohistochemical image based on yuv color space," in *Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on*, pp. 1–4, IEEE, 2010.

[46] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *Biomedical Engineering, IEEE Transactions on*, vol. 61, no. 6, pp. 1729–1738, 2014.

[47] T. Chen and C. Srinivasl, "Stain unmixing in brightfield multiplex immunohistochemistry images," in *MICCAI Machine Learning in Medical Imaging Workshop*, 2014.

[48] D. Carey, V. Wijayathunga, A. J. Bulpitt, and D. Treanor, "A novel approach for the colour deconvolution of multiple histological stains," in *Proceedings of the 19th Conference of Medical Image Understanding and Analysis*, pp. 156–162, BMVA, 2015.

[49] C. M. van der Loos, "Multiple immunoenzyme staining: methods and visualizations for the observation with spectral imaging," *Journal of Histochemistry & Cytochemistry*, vol. 56, no. 4, pp. 313–328, 2008.

[50] R. Montironi, L. Diamanti, D. Thompson, H. Bartels, and P. Bartels, "Analysis of the capillary architecture in the precursors of prostate cancer: recent findings and new concepts.," *European urology*, vol. 30, no. 2, pp. 191–200, 1995.

[51] A. C. Ruifrok, "Quantification of immunohistochemical staining by color translation and automated thresholding.," *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, vol. 19, no. 2, pp. 107–113, 1997.

[52] R. Rubin, D. S. Strayer, E. Rubin, *et al.*, *Rubin's pathology: clinicopathologic foundations of medicine*. Lippincott Williams & Wilkins, 2008.

[53] X. Chen, X. Zhou, and S. T. Wong, "Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 4, pp. 762–766, 2006.

[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[55] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 1491–1498, IEEE, 2006.

[56] F. Han, Y. Shan, R. Cekander, H. S. Sawhney, and R. Kumar, "A two-stage approach to people and vehicle detection with hog-based svm," in *Performance Metrics for Intelligent Systems 2006 Workshop*, pp. 133–140, 2006.

[57] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[58] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[59] P. Hobson, B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem, "Computer aided diagnosis for anti-nuclear antibodies hep-2 images: progress and challenges," *Pattern Recognition Letters*, vol. 82, pp. 3–11, 2016.

[60] I. Ersoy, F. Bunyak, J. Peng, and K. Palaniappan, "Hep-2 cell classification in iif images using shareboost," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3362–3365, IEEE, 2012.

[61] S. Ghosh and V. Chaudhary, "Feature analysis for automatic classification of hep-2 florescence patterns: Computer-aided diagnosis of auto-immune diseases," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 174–177, IEEE, 2012.

[62] R. Nosaka and K. Fukui, "Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns," *Pattern Recognition*, vol. 47, no. 7, pp. 2428–2436, 2014.

[63] R. Nosaka, Y. Ohkawa, and K. Fukui, "Feature extraction based on co-occurrence of adjacent local binary patterns," in *Pacific-Rim Symposium on Image and Video Technology*, pp. 82–91, Springer, 2011.

[64] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Hep-2 cells classification via sparse representation of textural features fused into dissimilarity space," *Pattern Recognition*, vol. 47, no. 7, pp. 2367–2378, 2014.

[65] X. Qi, G. Zhao, C.-G. Li, J. Guo, and M. Pietikäinen, "Hep-2 cell classification via fusing texture and shape information," *arXiv preprint arXiv:1502.04658*, 2015.

[66] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in*

computer vision, ECCV, vol. 1, pp. 1–2, Prague, 2004.

[67] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2, pp. 2169–2178, IEEE, 2006.

[68] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[69] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Computer Vision–ECCV 2006*, pp. 589–600, 2006.

[70] M. T. Law, N. Thome, and M. Cord, "Bag-of-words image representation: Key ideas and further insight," in *Fusion in Computer Vision*, pp. 29–52, Springer, 2014.

[71] L. Shen, J. Lin, S. Wu, and S. Yu, "Hep-2 image classification using intensity order pooling based features and bag of words," *Pattern Recognition*, vol. 47, no. 7, pp. 2419–2427, 2014.

[72] X.-H. Han and Y.-W. Chen, "Fisher vector of micro-texton for hep-2 staining pattern classification," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 3575–3580, 2014.

[73] S. Di Cataldo, A. Bottino, I. U. Islam, T. F. Vieira, and E. Ficarra, "Subclass discriminant analysis of morphological and textural features for hep-2 staining pattern classification," *Pattern Recognition*, vol. 47, no. 7, pp. 2389–2399, 2014.

[74] O. Sertel, U. V. Catalyurek, H. Shimada, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis

cells in digitized histological images," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 1433–1436, IEEE, 2009.

[75] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[76] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[77] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[78] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[82] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 411–418, Springer, 2013.

[83] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: Deep contour-aware networks for accurate gland segmentation," *arXiv preprint*

arXiv:1604.02677, 2016.

[84] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[85] Z. Gao, J. Zhang, L. Zhou, and L. Wang, "Hep-2 cell image classification with convolutional neural networks," in *Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 2014 1st Workshop on*, pp. 24–28, IEEE, 2014.

[86] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "Hep-2 cell image classification with deep convolutional neural networks.," *IEEE journal of biomedical and health informatics*, 2016.

[87] H. T. H. Phan, A. Kumar, J. Kim, and D. Feng, "Transfer learning of a convolutional neural network for hep-2 cell image classification," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pp. 1208–1211, IEEE, 2016.

[88] H. Li, J. Zhang, and W.-S. Zheng, "Deep cnns for hep-2 cells classification: A cross-specimen analysis," *arXiv preprint arXiv:1604.05816*, 2016.

[89] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Human epithelial type 2 cell classification with convolutional neural networks," in *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, pp. 1–6, IEEE, 2015.

[90] Y. Li and L. Shen, "A deep residual inception network for hep-2 cell classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 12–20, Springer, 2017.

[91] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, “Improved automatic detection and segmentation of cell nuclei in histopathology images,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[92] H. Fu, G. Qiu, J. Shu, and M. Ilyas, “A novel polar space random field model for the detection of glandular structures,” *IEEE transactions on medical imaging*, vol. 33, no. 3, pp. 764–776, 2014.

[93] A. Basavanhally, E. Yu, J. Xu, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, “Incorporating domain knowledge for tubule detection in breast histopathology using ocallaghan neighborhoods,” in *SPIE Medical Imaging*, vol. 7963, p. 796310, International Society for Optics and Photonics, 2011.

[94] S. K. Apple, “Sentinel lymph node in breast cancer: review article from a pathologists point of view,” *Journal of pathology and translational medicine*, vol. 50, no. 2, p. 83, 2016.

[95] I. Stage and I. Stage III, “Clinical oncology for students,”

[96] P. Shi, J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen, “Automated ki-67 quantification of immunohistochemical staining image of human nasopharyngeal carcinoma xenografts,” *Scientific reports*, vol. 6, p. 32127, 2016.

[97] Y. Li, L. Shen, and S. Yu, “Hep-2 specimen image segmentation and classification using very deep fully convolutional network,” *IEEE Transactions on Medical Imaging*, 2017.

[98] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, “Aggnet: deep learning from crowds for mitosis detection in

breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[99] 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.

[100] 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.

[101] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[102] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, 2017.

[103] P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478–486, Springer, 2016.

[104] G. Amit, O. Hadad, S. Alpert, T. Tlusty, Y. Gur, R. Ben-Ari, and S. Hashoul, "Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 594–602, Springer, 2017.

[105] M. Shah, C. Rubadue, D. Suster, and D. Wang, "Deep learning assessment of tumor proliferation in breast cancer histological images," *arXiv preprint arXiv:1610.03467*, 2016.

[106] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[107] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[108] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*, pp. 565–571, IEEE, 2016.

[109] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[110] E. Rakha, D. Soria, A. R. Green, C. Lemetre, D. G. Powe, C. C. Nolan, J. M. Garibaldi, G. Ball, and I. O. Ellis, "Nottingham prognostic index plus (npi+): a modern clinical decision making tool in breast cancer," *British journal of cancer*, vol. 110, no. 7, pp. 1688–1697, 2014.

[111] A. C. Ruifrok, R. L. Katz, and D. A. Johnston, "Comparison of quantification of histochemical staining by hue-saturation-intensity (hsi) transformation and color-deconvolution," *Applied Immunohistochemistry & Molecular Morphology*, vol. 11, no. 1, pp. 85–91, 2003.

[112] J. Shu, G. Qiu, M. Ilyas, and P. Kaye, "Biomarker detection in whole slide

imaging based on statistical color models," in *MICCAI 2010 Workshop on Computational Imaging Biomarkers for Tumors: From Qualitative to Quantitative*, 2010.

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

[114] H. Fu, Q. Zhang, and G. Qiu, "Random forest for image annotation," in *Computer Vision–ECCV 2012*, pp. 86–99, Springer, 2012.

[115] A. Kavanaugh, R. Tomar, J. Reveille, D. Solomon, and H. Homburger, "Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. american college of pathologists.," *Archives of pathology & laboratory medicine*, vol. 124, no. 1, pp. 71–81, 2000.

[116] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Hep-2 cells classification via fusion of morphological and textural features," in *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, pp. 689–694, IEEE, 2012.

[117] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Medical image analysis*, vol. 14, no. 1, pp. 1–12, 2010.

[118] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 2169–2178, 2006.

[119] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Computer Vision, 2005.*

ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp. 1458–1465, IEEE, 2005.

[120] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2911–2918, June 2012.

[121] G. V. Ponomarev, V. L. Arlazarov, M. S. Gelfand, and M. D. Kazanov, "{ANA} hep-2 cells image classification using number, size, shape and localization of targeted cell regions," *Pattern Recognition*, vol. 47, no. 7, pp. 2360 – 2366, 2014.

[122] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, pp. 2654–2662, 2014.

[123] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.

[124] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*, pp. 597–613, Springer, 2016.

[125] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[126] 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 *OSDI*, vol. 16, pp. 265–283, 2016.

[127] J. Liu, B. Xu, L. Shen, J. Garibaldi, and G. Qiu, "Hep-2 cell classification based on a deep autoencoding-classification convolutional neural

network," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pp. 1019–1023, IEEE, 2017.

[128] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "Hep-2 cell image classification with deep convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 21, no. 2, pp. 416–428, 2017.

[129] Y. Li and L. Shen, "A deep residual inception network for hep-2 cell classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 12–20, Springer, 2017.

[130] C. M. Perou, S. S. Jeffrey, M. Van De Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, *et al.*, "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *Proceedings of the National Academy of Sciences*, vol. 96, no. 16, pp. 9212–9217, 1999.

[131] T. O. Nielsen, J. S. Parker, S. Leung, D. Voduc, M. Ebbert, T. Vickery, S. R. Davies, J. Snider, I. J. Stijleman, J. Reed, *et al.*, "A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer," *Clinical cancer research*, pp. 1078–0432, 2010.

[132] D. Soria, J. M. Garibaldi, F. Ambrogi, A. R. Green, D. Powe, E. Rakha, R. D. Macmillan, R. W. Blamey, G. Ball, P. J. Lisboa, *et al.*, "A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients," *Computers in biology and medicine*, vol. 40, no. 3, pp. 318–330, 2010.

[133] A. Green, D. Powe, E. Rakha, D. Soria, C. Lemetre, C. Nolan, F. Barros, R. Macmillan, J. Garibaldi, G. Ball, *et al.*, "Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers," *British journal of cancer*, vol. 109, no. 7, p. 1886, 2013.

[134] H. Goulding, S. Pinder, P. Cannon, D. Pearson, R. Nicholson, D. Snead, J. Bell, C. Elston, J. Robertson, R. Blamey, *et al.*, "A new immunohistochemical antibody for the assessment of estrogen receptor status on routine formalin-fixed tissue samples," *Human pathology*, vol. 26, no. 3, pp. 291–294, 1995.

[135] P. Haub and T. Meckel, "A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration," *Scientific reports*, vol. 5, 2015.

[136] P. Hobson, B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem, "Hep-2 staining pattern recognition at cell and specimen levels: datasets, algorithms and results," *Pattern Recognition Letters*, vol. 82, pp. 12–22, 2016.

[137] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," 2017.

[138] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, 2017.

[139] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[140] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[141] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nih image to imagej: 25 years of image analysis," *Nature methods*, vol. 9, no. 7, pp. 671–675, 2012.

[142] J. Shu, H. Fu, G. Qiu, P. Kaye, and M. Ilyas, "Segmenting overlapping cell nuclei in digital histopathology images," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 5445–5448, IEEE, 2013.

[143] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: context, process, and purpose," 2016.

[144] S. Kothari, J. H. Phan, and M. D. Wang, "Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade," *Journal of pathology informatics*, vol. 4, 2013.