Deep Representation Learning of High-Resolution Whole Slide Histopathology Images

by

Rasoul Sali

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
at the
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA

March 2021

Committee Approval:

We, the undersigned committee members, certify that we have reviewed this dissertation and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Systems and Information Engineering.

Donald E. Brown, Ph.D. (School of Engineering and Applied Science) Date
Advisor

Michael D. Porter, Ph.D. (School of Engineering and Applied Science) Date
Committee Chair

Laura Barnes, Ph.D. (School of Engineering and Applied Science) Date
Committee Member

Sana Syed, MD, MS, MSDS (School of Medicine) Date
Committee Member

Scott Schwartz, Ph.D. (School of Data Science) Date
Committee Member

Certified by:

Craig H. Benson, Dean, School of Engineering and Applied Science Date

©[2021] [Rasoul Sali]
All rights reserved.

Abstract

Whole slide tissue histopathology images (WSIs) play a crucial role in tissue specimen assessment and diagnosis of associated diseases. Recent technological progress in image acquisition systems has led to an increasing accumulation of high-resolution histopathology images. Nevertheless, employing these images to develop clinical decision support systems has been hampered by the need for manual examination of WSIs, a subjective, labor-intensive, time-consuming, and error-prone process. This creates a burgeoning demand for new analytic approaches to analyze/pre-process such images. Furthermore, a content-based representation allows the integrated study of histopathology images with other data modalities enabling holistic and multi-modal analysis of human diseases. Deep learning approaches have shown promising performance on feature extraction from images. However, dealing with WSIs introduces new challenges, demanding more efficient approaches to learn an informative representation of these images. This research aims to employ deep learning approaches for representation learning of WSIs focusing on Barrett's Esophagus (BE). In this setting, three different approaches will be considered: Bag of Visual Words (BoVW), Neural Image Compression (NIC), and Graph Neural Networks (GNN).

Index terms— Computational pathology, Deep learning, Representation learning, Whole-slide histopathology images, Bag of visual words, Neural image compression, Graph neural networks, Barrett's esophagus

Acknowledgments

I would like to express my sincere gratitude to my esteemed advisor, Prof. Donald Brown, for providing me with an extremely interesting project and his invaluable supervision, continuous support, and patience during my Ph.D. career. Additionally, I would like to extend my gratitude and thankfulness to Dr. Sana Syed for her treasured support, which was really influential in critiquing my results. I would also like to thank Dr. Michael Porter, Dr. Laura Barnes, and Dr. Scott Schwartz, my dissertation advisory committee members, whose supports and comments helped me improve this thesis.

Without the love and support of my family, this would have been an arduous journey. I would like to thank my parents for their unconditional support and constant encouragement to pursue my goals. I also want to thank my kids for their cheering and charming inspirations. Finally, I would like to thank and dedicate this work to my wife for giving me unwavering love and support and also for being so patient with me and putting up with my long hours of research.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BE	Barrett's Esophagus
BiGAN	Biderectional Generative Adversarial Network
BoVW	Bag of Visual Word
CAE	Convolutional Autoencoder
CBGCN	Cluster-Based Graph Convolution Network
CBIR	Content-Based Image Retrieval
CBNIC	Cluster-Based Neural Image Compression
CD	Celiac Disease
CNN	Convolutional Neural Network
DSIFT	Dense Scale Invariant Feature Transform
GCN	Graph Convolution Network
GI	Gastrointestinal
GNN	Graph Neural Network
H&E	Hematoxylin & Eosin
HD-WLE	High Definition White-Light Endoscopy
LBP	Local Binary Patterns
MIL	Multiple Instance Learning
NBI	Narrow-Band Imaging
NIC	Neural Image Compression
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
RoI	Region of Interest
SIFT	Scale Invariant Feature Transform
SURF	Speeded-Up Robust Features

SVM	Support Vector Machine
TF	Term Frequency
WSI	Whole-Slide histopathology Image

1 Introduction

Barrett's esophagus (BE) is a potentially severe condition that results from damage to the lining of the squamous esophageal mucosa because of gastroesophageal reflux disease. Its diagnosis is based on the endoscopic and histologic findings of the columnar epithelium lining the distal esophagus [2]. To increase sensitivity for dysplasia, guidelines recommend the Seattle protocol, which involves taking four-quadrant random biopsies at 1–2 cm intervals [3]. However, this protocol does not permit real-time diagnosis or therapy and is labor-intensive, leading to low adherence [4, 5]. Furthermore, numerous studies have documented inter-observer variability among pathologists when diagnosing both low-grade [6,7] and high-grade dysplasia [8], which are the stages through which BE progresses before becoming esophageal cancer. Because the diagnosis of dysplastic and non-dysplastic BE can improve clinical care and prevent disease complications, there is a clear need for an accurate diagnostic tool that translates heterogeneous histopathology images into accurate and precise diagnostics. The development of such a system in high-dimensional clinical research will support precision medicine with improved diagnostics, predictions, treatments, and patient clinical outcomes. The success of these systems relies on how well they extract morphological image features and characterize the images' visual content.

Whole slide tissue histopathology images (WSIs) are the gold standard for diagnosing the presence, type, and progression of several diseases, including the most type of cancer [9] and also some Gastrointestinal (GI) disorders including BE [10, 11]. WSIs being rich in information and preserving the underlying tissue structure, provide a comprehensive view of diseases and their effect on the tissue [9]. The advent of WSI scanners has made possible the virtualizing and digitalizing of the whole glass slides [12]. This has led to an increased accumulation of digital histopathology images. Thus, their utilization for clinical decision-making needs to keep pace with the rise in their digitization rate. However, this has been hampered by the need for manual examination of WSIs, a subjective, labor-intensive, time-consuming, and error-prone process [13]. It creates a burgeoning need for developing new analytic approaches for the automated analysis of histopathology images. The extracted features from histopathology images can be combined with other data modalities and multi-

omics data to provide clinicians with more precise diagnostic, prognostic, and therapeutic determinations.

Our main contribution in this dissertation is proposing models for representation learning of esophageal WSIs. These deep learning based models encode WSIs using visual words while capturing the spatial proximity information between local features and finally provide an image-wise representation accordingly. At the same time, only the reported diagnoses as image labels have been utilized for training. We provide experimental evidence that employing visual words rather than patch-level representation vectors might be beneficial for the classification of WSIs. Furthermore, the interpretability of models is improved because the proposed models deal with a dictionary containing a finite number of visual words. These models can be part of a clinical decision support system to assist practitioners and pathologists in image-level interpretation tasks and diagnosis of different classes of BE.

In this dissertation, three different approaches will be applied for representation learning of esophageal WSIs: Bag of Visual Words (BoVW), Cluster-Based Neural Image Compression (CBNIC), and Cluster-Based Graph Convolutional Network (CBGCN). Each model has two main steps: in the first step, unsupervised deep feature extraction approaches (e.g., Convolutional Autoencoder (CAE) and Bidirectional Generative Adversarial Network (BiGAN)) are exploited to extract local tissue-derived image features. The main reason for applying an unsupervised approach is to avoid expensive and time-consuming image annotation. Furthermore, employing an unsupervised approach makes it possible to detect disease-associated image patterns unknown to pathologist annotators, especially in the case of diseases whose etiology has not yet been well explored. In the second step, local image features are aggregated to provide an image-level representation. The first step is the same in all models, but the models introduce different aggregation mechanisms. The effectiveness of each representation learning approach will be evaluated in a supervised paradigm for the classification and localization of WSIs.

The remainder of this document is organized as follows: Chapter 2 briefly reviews the literature on the analysis of WSIs. The dataset was employed in this dissertation, and local feature quantization approaches are presented in Chapter 3. Chapters 4, 5, and 6 present the BoVW, CBNIC, and CBGCN for representation learning of WSIs respectively. Finally, Chapter 7 concludes this dissertation along with outlining future directions.

2 | Literature Review

There are some problems associated with conventional feature engineering approaches. First, the combinatorial nature of the feature extraction process makes it expensive to hand-craft features. Furthermore, the development of these features commonly relies on task/domain-specific expertise, preventing them from adapting to new tasks or domains. Also, human bias is an inseparable part of hand-crafted features. In recent years, deep learning approaches have revolutionized the process of feature extraction tasks in the computer vision domain, among others [14]. However, dealing with WSIs arises new challenges, demanding more effective approaches to learn the representation of these images.

Some of the challenges regarding the analysis of whole-slide histopathology images are as follows. First, WSIs are gigapixel images (typically $100,000 \times 100,000$ RGB pixels), and given the current technology, it is infeasible to directly train a CNN model on such images due to steep computational requirements [9, 15, 16]. Also, image down-sampling as a solution to this problem leads to the informative details loss at cell level [16]. Second, in most cases, only the image-wise ground truth label is given because of the high cost of pixel-wise annotation on high-resolution images [16]. Third, histopathology images are heterogeneous and contain a large amount of biologically related spatial variation [9]. Furthermore, the overall shape of the target regions might differ significantly from one image to another, which makes their pattern difficult to learn, and hence to locate. Moreover, regions of interest typically have visual appearances that are pretty similar to the surrounding tissues and normal regions, making them much more difficult to distinguish from the background. Fourth, while staining is crucial as it enables visualization of the microscopic structural features in the biopsy, variation in the H&E staining process across different lab sites can lead to variations in biopsy image appearance [9, 17, 18]. These variations introduce an undesirable bias when the slides are used to train machine learning models.

Feature acquisition from high-resolution tissue tiles sampled from WSIs is considered a potential solution to directly address some of the hurdles mentioned above and provide a base for better solving other issues. In this approach, the WSI-wise representation is acquired by combining the local image features extracted from sampled tissue tiles [19, 20, 21, 22, 23].

There is a rich body of literature investigating feature representation in the form of three primary approaches: fully-supervised [24,25,26], weakly supervised [9,11,12,27,28,29], and unsupervised feature learning [30,31,32,33,34].

Of these, the fully supervised feature learning approach requires a large amount of accurately annotated data, which can be a labor-intensive, time-consuming, and error-prone process. These challenges are abundantly clear in the classification and segmentation of histopathology images, as accurate and complete annotations can be difficult even for expert pathologists. On the opposite end of the annotation spectrum, unsupervised approaches aim to learn a discriminative representation of WSIs from annotation-free histopathology images. These methods extract the salient features from WSIs without requiring any image-level diagnosis as an image label or region of interest annotated by experts. Finally, weakly supervised methods have the advantages of both the fully supervised and unsupervised approaches for feature learning [35]. Weakly supervised techniques exploit coarsely grained annotated WSIs to simultaneously classify histology images and yield pixel-wise localization scores, thereby identifying the corresponding regions of interest.

After extracting the patch-wise feature representations, they are aggregated to provide an image-level representation. In the literature, different aggregation approaches have been proposed including bag of visual words [31, 36, 37, 38, 39], graph neural networks [40,41,42,43,44], neural image compression [15,45], attention-based neural networks [46,47], etc.

3 | Local Features Extraction

3.1 Esophageal Biopsies Dataset

This study utilizes previously published preliminary data to apply deep learning techniques for detecting BE and dysplasia in Hematoxylin and Eosin (H&E) stained biopsies. All patients in the study conducted by Shah et al. [48] (years 2014–2016) underwent targeted biopsy or mucosal resection and Seattle protocol biopsies. To increase the sample size, a retrospective chart review was conducted to identify and retrieve biopsy slides of patients who had undergone upper endoscopies for BE surveillance (years 2016–2019). These patients all underwent high-definition white-light endoscopy (HD-WLE), narrow-band imaging (NBI), and acetic acid chromoendoscopy followed by targeted biopsies/mucosal resection, and Seattle protocol biopsies. All biopsy specimens were fixed in formalin. Samples were embedded to exhibit the full mucosal thickness. The paraffin blocks were sectioned at three microns to create biopsy slides that were stained with hematoxylin and eosin. All suspected diagnoses of dysplasia or malignancy required a consensus of two or more pathologists. For patients included in Shah et al.'s study [48], a blinded expert pathologist also reviewed all biopsy specimens. Blinded and unblinded pathology results were prospectively recorded.

This study was approved by the Hunter Holmes McGuire Veterans Affairs Medical Center Institutional Review Board and the University of Virginia Institutional Review Board for Health Science Research (IRB-HSR #21328).

Tissue images were digitized at 40× magnification via scanning of biopsy slides using a Hamamatsu NanoZoomer S360 Digital slide scanner C13220 [49]. A total of 387 whole-slide images from 133 unique patients were collected. WSIs increased to 650 after pre-processing and cropping; 115 whole-slide images from 13 patients were selected to train deep models to extract patch-level image features in all three feature learning approaches, and the rest of the dataset was used for model evaluation. To train deep models in fully supervised approaches, these WSIs were manually pixel-wise annotated to highlight each class's examples within each whole-slide image (see Figure 3.1).

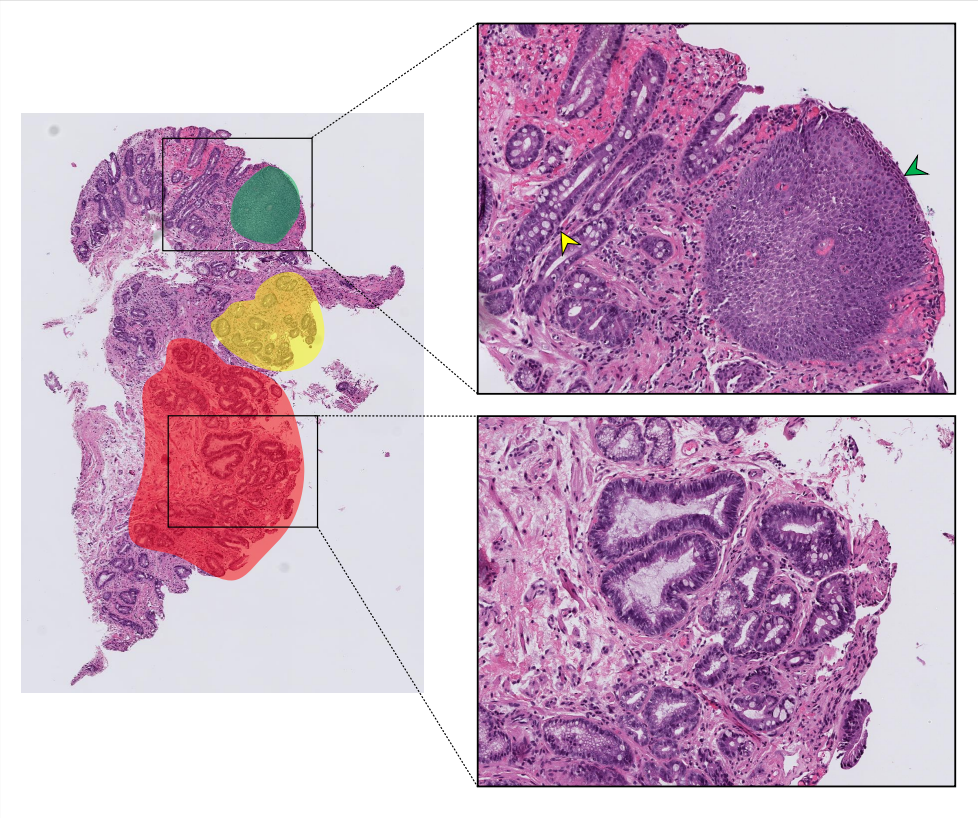

Figure 3.1: An example of the annotation process on a typical whole-slide image (WSI). Red, green, and yellow highlighted areas indicate areas that were annotated and from which labeled patches were taken. Squamous tissue (green arrowhead), non-dysplastic Barrett's with Goblet cells (yellow arrowhead), and dysplastic tissue with crowding and hyperchromasia (lower zoomed section) were all present within the same whole-slide image.

3.2 Data Preparation

It is worth noting that the same dataset is used to evaluate different approaches in this dissertation. Since the final goal is evaluating our models on WSIs, each WSI constitutes a single training data point. In this setup, our dataset consists of only a few hundred WSIs, and the risk of over-fitting is considerable when training a deep model with millions of parameters. Furthermore, training a CNN requires images of the same size; however, usually, WSIs, even in a single dataset, have different sizes. Resizing these images as a solution to this issue arises some other hurdles such as resolution and scale variation.

To have enough images of the same size for CNN training, we first employed a sliding window method on each WSI at $40\times$ magnification to generate very large patches of size 5000×5000 pixels. We assume that each large patches have all image features of the original corresponding WSI to consider them as new WSIs with the same label as the original ones.

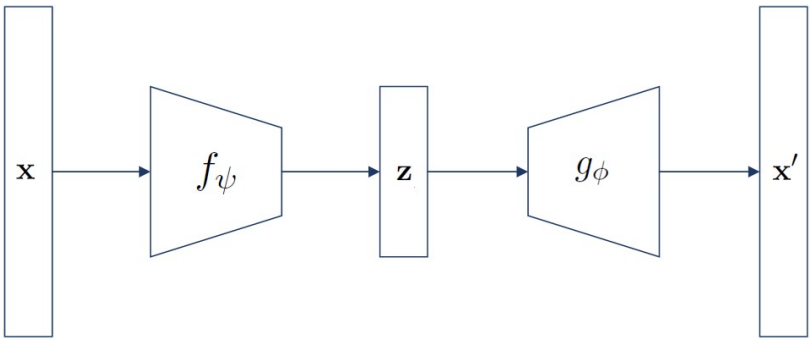

Figure 3.2: The structue of Convolutional Autoencoder (CAE)

Large patches with less than 50% tissue sections were discarded. Image augmentation was also performed by horizontal flipping, random 90-degree rotations, and image mirroring during training to prevent CNNs from over-fitting. Then, we again employed the sliding window method, this time on the large patch at 40× magnification to generate tissue tiles of size 200×200 pixels. The tissue tiles were resized to 128×128 pixels.

A common issue that causes bias while training the model on histopathological images is color variation. This issue, which originates from various sources, including differences in raw materials, staining protocols, and digital scanners [18], should be addressed and resolved as an essential pre-processing step before any analyses. Various solutions, such as color balancing [50], gray-scale, and stain normalization, have been proposed in the published literature to address the color variation issue. In this study, we used Deep Convolutional Gaussian Mixture Model (DCGMM) [51] to address the color variation issue.

3.3 Quantization of Local Image Features

We consider CAE and BiGAN to extract tile-level features.

3.3.1 Convolutional Autoencoder

An autoencoder (see Figure 3.2) is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation for a set of data, typically for dimensionality reduction [52]. Generally speaking, autoencoders consists of two parts; encoder and decoder. Encoder $f(.)$ is a function parameterized by ψ that maps input $\mathbf{x}$ to a hidden space $\mathbf{z}$; $\mathbf{z} = f_\psi(\mathbf{x})$ and a decoder is function $g(.)$ parameterized by ϕ that produces a reconstruction $\mathbf{x}' = g_\phi(\mathbf{z})$. The optimal weights for encoder and decoder are derived from equation 3.1.

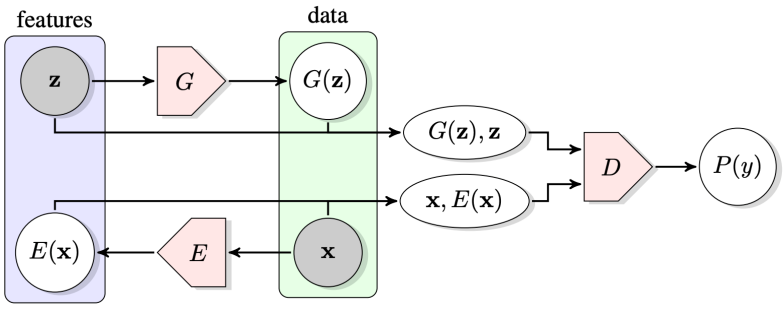

Figure 3.3: The structue of Bidirectional Generative Adversarial Network (BiGAN) [1]

$$f^*, g^* = \arg\min_{\psi,\phi} \|\mathbf{x} - g_\phi(f_\psi(\mathbf{x}))\|^2 = \arg\min_{\psi,\phi} \|\mathbf{x} - \mathbf{x}'\|^2 \tag{3.1}$$

3.3.2 Bidirectional Generative Adversarial Network

As shown in the Figure 3.3 the BiGAN [1, 53] consists of three networks: a generator G, which maps a latent variable $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to generated images $\mathbf{x}'$:

$$\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \xrightarrow{G} \mathbf{x}' = G(\mathbf{z}) \sim p(\mathbf{x} \mid \mathbf{z}) \tag{3.2}$$

an encoder E, which maps each image $\mathbf{x}$ sampled from the empirical data distribution $q(.)$ to embedding space $\mathbf{z}'$:

$$\mathbf{x} \sim q(\mathbf{x}) \xrightarrow{E} \mathbf{z}' = E(\mathbf{x}) \sim q(\mathbf{z} \mid \mathbf{x}) \tag{3.3}$$

and a discriminator D which during an adversarial minimax game is trained to discriminate joint samples of the data and the corresponding latent variable from the encoder $(\mathbf{x}, E(\mathbf{x}))$ from joint samples of the generator $(G(\mathbf{z}), \mathbf{z})$ that are drawn form $q(\mathbf{x}, \mathbf{z})$ and $p(\mathbf{x}, \mathbf{z})$ respectively. While encoder and generator networks are trained to fool the discriminator. The optimal functions are derived from equation 3.4.

$$\begin{aligned} E^*, G^*, D^* = & \arg\min_{E,G} \max_{D} \{\log D(\mathbf{x}, E(\mathbf{x})) + \log(1 - D(G(\mathbf{z}), \mathbf{z}))\} \\ = & \arg\min_{E,G} \max_{D} \{\log D(\mathbf{x}, \mathbf{z}') + \log(1 - D(\mathbf{x}', \mathbf{z}))\} \end{aligned} \tag{3.4}$$

3.4 Architecture of Feature Extractors

In CAE architecture, ResNet18 was employed as an encoder. We removed fully connected layers from the original network and employed the ResNet backbone as a feature extractor, followed by a dense layer that received the flattened output of the feature extractor. The decoder comprised convolutional and up-sampling layers to increase the size of the feature maps and get back the original size of the input image. The size of embedding space is a critical parameter. When the embedded layer's size is large, the network is not forced to learn informative features. On the other hand, selecting a small value for the latent space size makes it impossible for the model to reconstruct the input images. In this study, we set the size of embedding layer to 256.

In BiGAN, the encoder consists of six 2D convolutional layers, which maps each tissue tile to an embedding vector size of 256. The generator is composed of 6 deconvolutional layers to map input signals into generated image space. In both networks, the rectified linear unit (ReLU) [14] was employed as the activation function.

4 | Bag of Visual Words

4.1 Background

The Bag of Visual Words (BoVW) approach or dictionary learning was inspired by the bag-of-words scheme proposed initially for text categorization and text retrieval [31]. In this setting, an image is treated as a document, and the image features as words. Through this method, local image features are extracted and quantified to construct a visual words dictionary (i.e., the visual codebook). Finally, an image is encoded as an order-less histogram of visual word frequencies. The informativeness of image-level representation is investigated in a supervised fashion.

The BoVW approach has been widely used in the medical image domain for image annotation, classification, and retrieval and has shown a solid performance [38]. Avni et al. [54] used BoVW to encode SIFT descriptors for categorization of chest X-ray images and achieved the top performance in ImageCLEF competition for medical image annotation task, which is based on the IRMA project X-ray library [55]. Powell et al. [56] applied BoVW approach on histopathology images to extract tissue-derived image features and used them to predict the overall survival in lower grade Gliomas. A support vector machine (SVM) model was applied to discriminate patients into short and long overall survival groups dichotomized at 24-month. Bardou et al. [57] used BoVW to encode local histology descriptors extracted by Dense Scale Invariant Feature Transform (DSIFT) features and Speeded-Up Robust Features (SURF) for the classification of breast cancer. Zhang et al. [58] proposed a BoVW scheme using the sparse random feature to classify epithelial nuclei, and stroma nuclei objects to segment the glandular structure in histology images of colon tissues. Mittal et al. [59] applied BoVW and gravitational search algorithm to encode histological contents to classify the images into the respective tissue categories to facilitate the quantification analysis of histopathology images by removing inter-category heterogeneity. Yamamoto et al. [30] used the BoVW approach on histological image features extracted by a deep Convolutional Autoencoder (CAE) in an unsupervised fashion to encode histology images to predict prostate cancer recurrence.

4.2 Model Development

The general idea of the BoVW approach is representing an image as an orderless set of visual features. The BoVW consists of two main steps: In the first step, a visual codebook is learned for representing the images of interest. A codebook is a visual vocabulary $V = \{c_1, ..., c_K\}$ including representative local descriptors codified as K visual words. In the second step, each local image feature is associated with a visual word. Finally, the image is represented by the histogram of the codeblocks [1]. Figure 4.1 represents the overview of BoVW model. The main steps of this model are as follows: image encoding and training a classifier on image-level histograms.

4.2.1 Image Encoding

Codebook Learning

In this step, a visual codebook is constructed. After dividing the gigapixel image w into a set of high-resolution tissue tiles $x_{ij} \in \mathbb{R}^{P \times P \times 3}$ sampled from the i^{th} row and j^{th} column of an uniform grid of square patches of size P using a stride of S throughout w, each tissue tile x is mapped into a low-dimensional embedding vector size E independently. This study employed CAE and BiGAN trained in an unsupervised fashion to map each high-resolution tissue tile into a low-dimensional embedding space. Then, a k-means clustering algorithm is employed to cluster extracted features into several visual words (see Fig 4.1, A-1). Selection of the number of clusters (codebook size) is an important decision in codebook construction. This parameter should be guessed/optimized and then imported to the model as an an input.

WSI Encoding

After learning the set of visual words, for WSI $X_i = \{x_1, ..., x_n\}$ including n tissue tile, the tile-wise embedding vectors are assigned to visual words (see Fig 4.1, A-2). Then, the histogram of codeblocks' frequencies $H_i = (h_1, ..., h_K)$ is considered as representation of X_i. The k-th bin in H_i is calculated as follows:

$$h_k = \frac{1}{|X_i|} \sum_{x \in X_i} p(c_k | f_\psi(x)) \qquad k = 1, ..., K. \tag{4.1}$$

where, $p(c_k|f_\psi(x))$ is the likelihood that embedding vector of tissue tile x (learned by neural network $f_\psi(.)$ with parameter ψ), belongs to codeblock c_k. As can be seen, the image-level histogram is the normalized frequency of each visual word c_k to relieve the effect of the

[1] in this dissertation, visual word, codeblock, and codeword are used interchangeably.

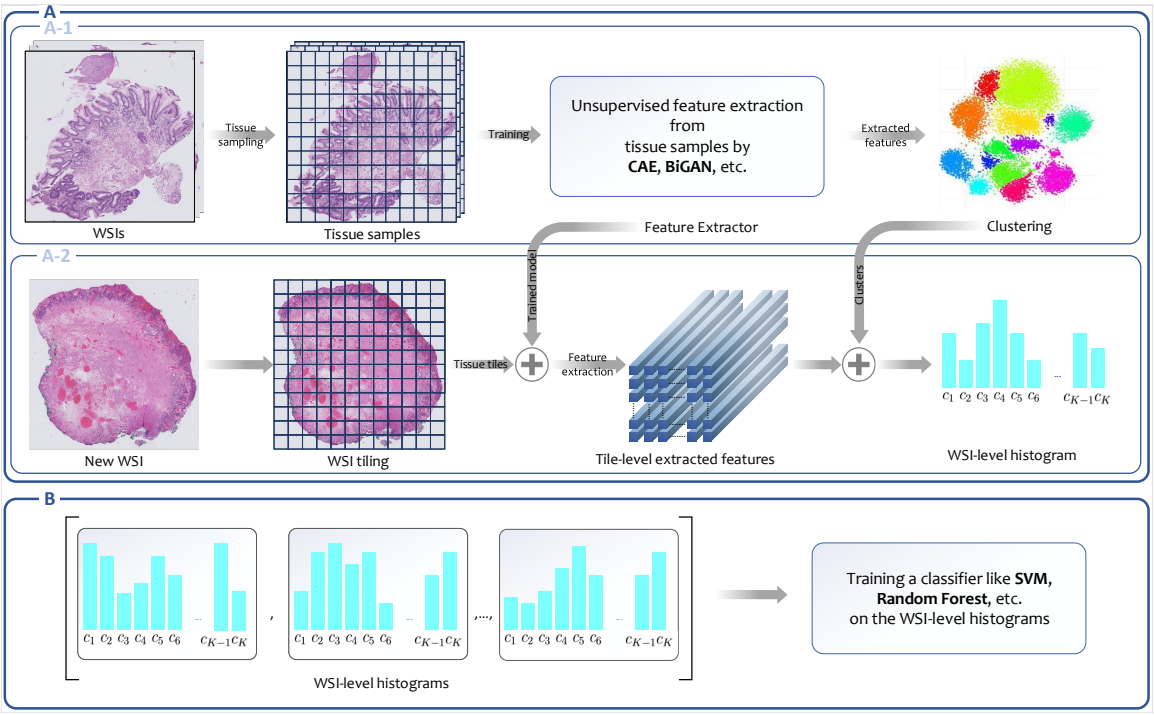

Figure 4.1: Overview of BoVW approach

number of tissue tiles in cases that WSIs have different number of tissue tiles. The hard assignment or soft assignment of patches to the clusters can be considered, depending on which clustering algorithm is used. In the case of employing k-means clustering, which gives a hard assignment of instances to clusters, image-level histogram values are calculated based on Equation (4.2).

$$h_k = \frac{1}{|X_i|} \sum_{x \in X_i} \mathbb{I}(c_x = k) \qquad k = 1, ..., K, \tag{4.2}$$

where h_k is the value of bin k-th in the generated histogram, and c_x is the cluster that image patch x belongs to.

4.2.2 Slide-Level Inference

After encoding WSIs, the image-level histograms are employed to train a classifier to predict the WSI-level labels.

4.2.3 Feature Importance

After training a classifier on encoded WSIs, the importance of codeblocks calculated by the model can be considered to evaluate the model's performance qualitatively. Per-class

importance of each tissue tile x for class C, $\mathcal{I}_x^C$ is calculated as follows:

$$\mathcal{I}_x^C = \sum_{m=1}^{K} p(c_m|f_\psi(x))I_{c_m}^C, \tag{4.3}$$

where $p(c_m|f_\psi(x))$ is the posterior probability of codeblock m-th given $f_\psi(x)$ (embedding vector of tissue tile x), and $I_{c_m}^C$ is the importance of the same codeblock for class C. We used the permutation feature importance to calculate per-class importance of each feature. In this method, each feature's importance for a specific class is defined to be the increase in the models' prediction error when values of that feature are randomly shuffled [60] which demolishes the relationship between the feature and the label. Model error increase due to shuffling the values of a feature shows that this feature is relevant. In this case, the model relied on the given feature for the prediction. In contrast, a feature is irrelevant if the model error does not change after shuffling the feature's values. In this case, the model ignores this feature for the prediction.

4.3 Experiments and Results

4.3.1 Experimental Results

The performance of a classification model is highly correlated with the degree of separability between different classes. Before applying classification algorithms on encoded WSIs, we visualized image-level representation vectors using the principal component analysis (PCA) method to understand better how well each method characterizes the visual content of histopathology images (see Figure 4.2). What can be deduced from the graphs is that both feature learning approaches have generated very similar results. In both models, lots of squamous WSIs were encoded relatively separately from dysplastic and non-dysplastic BE. However, as expected, there is confusion between these two classes.

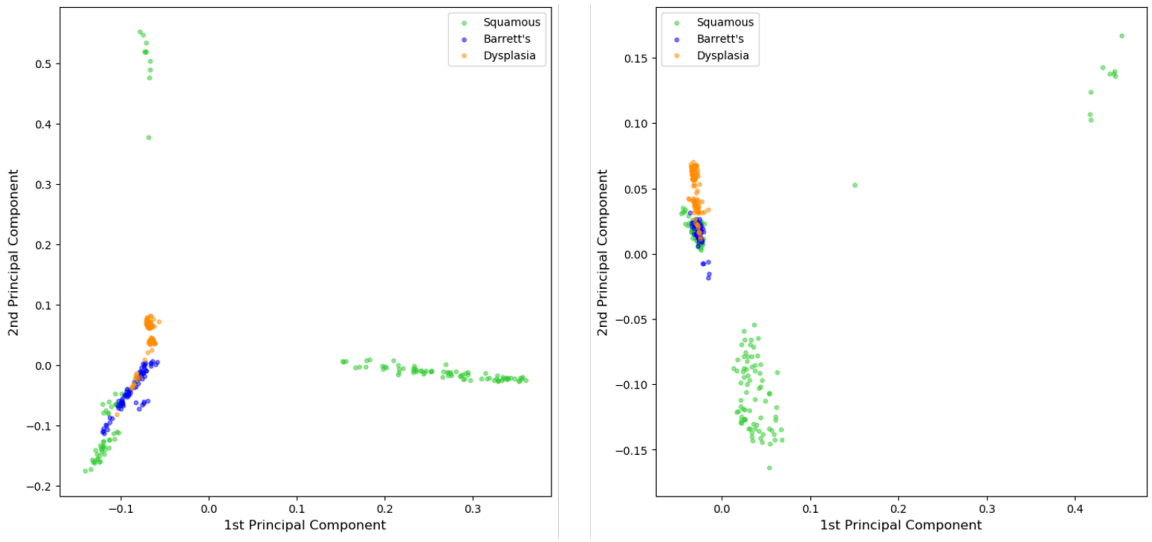

Figure 4.2: Principal component analysis (PCA) plot for WSIs encoded using (left) BoVW-CAE, (right) BoVW-BiGAN.

Classification results can further refine the findings from the PCA plots. Five standard metrics were used for classification under a 1-vs-rest strategy: accuracy, precision, recall, specificity, and F1 score. To estimate the significance of results, bootstrapping was used for all metrics.

After applying the sliding window method on 650 WSIs from 130 unique patients, a total of 2135 big patches (5000×5000 pixels) were generated, of which 793 (37.1%) were in the squamous class, 606 (28.4%) in non-dysplastic BE, and 736 (34.5%) in dysplastic BE. Of the independent testing set of 321 images, 142 (44.2%) squamous, 74 (23.1%) non-dysplastic BE, and 105 (32.7%) dysplastic BE images were used to evaluate trained models and to analyze the classification performance from both quantitative and qualitative aspects.

Results of esophageal WSIs classification in three classes of squamous, dysplastic BE, and non-dysplastic BE for different models are summarized in Table 4.1. The reported values are averages with 95% confidence intervals. For computing confidence intervals, numbers greater than one were truncated to 1.

Table 4.1: Results of WSI classification using BoVW approach

Class	Metric	Model	
		BoVW-CAE	**BoVW-BiGAN**
Squamous	Accuracy	0.841 (0.837, 0.845)	0.902 (0.899, 0.904)
	Precision	0.896 (0.890, 0.901)	1.000 (1.000, 1.000)
	Recall	0.725 (0.717, 0.732)	0.778 (0.772, 0.784)
	Specificity	0.933 (0.930, 0.937)	1.000 (1.000, 1.000)
	F1 score	0.801 (0.795, 0.806)	0.875 (0.871, 0.879)
Barrett's	Accuracy	0.842 (0.838, 0.846)	0.796 (0.792, 0.800)
	Precision	0.593 (0.584, 0.601)	0.533 (0.525, 0.541)
	Recall	1.000 (1.000, 1.000)	0.948 (0.943, 0.953)
	Specificity	0.794 (0.789, 0.799)	0.751 (0.745, 0.756)
	F1 score	0.743 (0.737, 0.750)	0.681 (0.674, 0.688)
Dysplasia	Accuracy	0.926 (0.923, 0.929)	0.894 (0.891, 0.898)
	Precision	1.000 (1.000, 1.000)	0.951 (0.946, 0.955)
	Recall	0.775 (0.767, 0.782)	0.713 (0.704, 0.721)
	Specificity	1.000 (1.000, 1.000)	0.982 (0.980, 0.984)
	F1 score	0.872 (0.868, 0.877)	0.814 (0.808, 0.820)
Weighted Average	Accuracy	0.869 (0.866, 0.872)	0.875 (0.872, 0.877)
	Precision	0.861 (0.859, 0.864)	0.877 (0.875, 0.879)
	Recall	0.804 (0.800, 0.809)	0.796 (0.792, 0.800)
	Specificity	0.923 (0.921, 0.925)	0.937 (0.935, 0.938)
	F1 score	0.812 (0.808, 0.816)	0.811 (0.807, 0.815)

As shown in Table 4.1, two models have similar performance given the weighted average of F1 score for all classes as an evaluation metric. The F1 score for the model trained on encoded WSIs using CAE is 0.812 (95% CI, 0.808-0.816) vs 0.811 (95% CI, 0.807-0.815) for the model trained on encoded WSIs using BiGAN. Tables 4.2 and 4.3 shows the confusion matrix of BoVW-CAE and BoVW-BiGAN respectively.

Table 4.2: Confusion matrix of BoVW-CAE

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	106 (0.746)	36 (0.254)	0 (0.000)
	Non-dysplastic BE	0 (0.000)	74 (1.000)	0 (0.000)
	Dysplastic BE	8 (0.076)	13 (0.124)	84 (0.800)

Table 4.3: Confusion matrix of BoVW-BiGAN

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	110 (0.775)	32 (0.225)	0 (0.000)
	Non-dysplastic BE	0 (0.000)	70 (0.946)	4 (0.054)
	Dysplastic BE	0 (0.000)	30 (0.286)	75 (0.714)

As the number of clusters is a critical parameter in the performance of clustering methods, we evaluated different numbers of clusters (codeblocks) for both approaches to pick a decent number of codeblocks given our dataset. Figure 4.3 summarizes the results

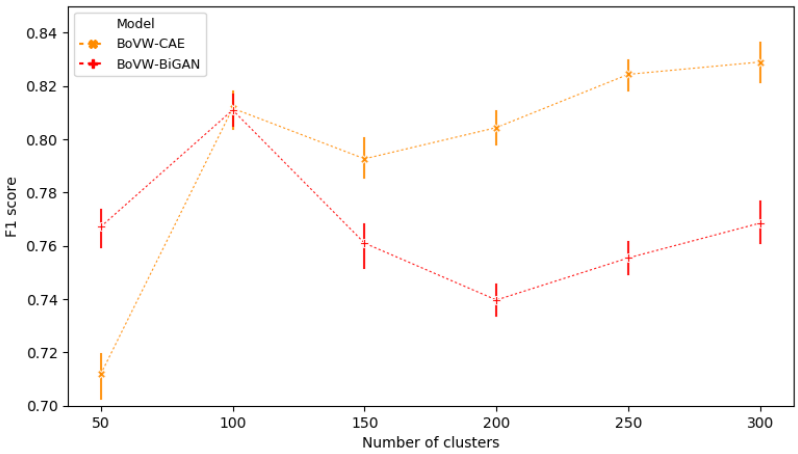

Figure 4.3: The effect of the number of clusters on the performance of the BoVW

of the evaluation of different numbers of clusters for both CAE and BiGAN approaches. As shown, the same trend can be seen for both approaches. After a significant increase in the model performance due to increasing the number of clusters from 50 to 100, the weighted F1 score decreases as a result of increasing the number of clusters, and again with further increase in the number of clusters, model performance is improved. Regarding the BoVW-CAE, although increasing the number of clusters improves the model performance, the complexity of the model also increases, leading to an undesirable consequence, a decrease in interpretability, which is crucial, especially in the field of medical image analysis. Thus, in cases that the improvement in model performance due to the increase in the number of clusters is not highly significant, the less number of clusters, the preferable model we will have. Thus, 100 is determined as an optimal value for the number of clusters, although 250 and 300 clusters have better classification results. Also, for BoVW-BiGAN, 100 is selected as the optimal number of clusters.

As can be seen, the performance of BoVW-CAE is better than BoVW-BiGAN for different number of clusters. Furthermore, compared to BoVW-CAE, BoVW-BiGAN is more sensitive to number of clusters, and selecting an optimal number of clusters is more crucial in this model.

4.3.2 Analysis of Codeblocks

For more scrutiny about the models' performances, the per-class importance of each visual word that is a measure of how important it is in the image-level inference is calculated by Equation 4.3. By aggregating these values from all images in the test set for every visual word and averaging, a value is obtained indicating the relative importance of given codeblock. Figures 4.4 and 4.5 show some randomly selected image patches associated with the top five

visual words in each class for both BoVW-CAE and BoVW-BiGAN, respectively.

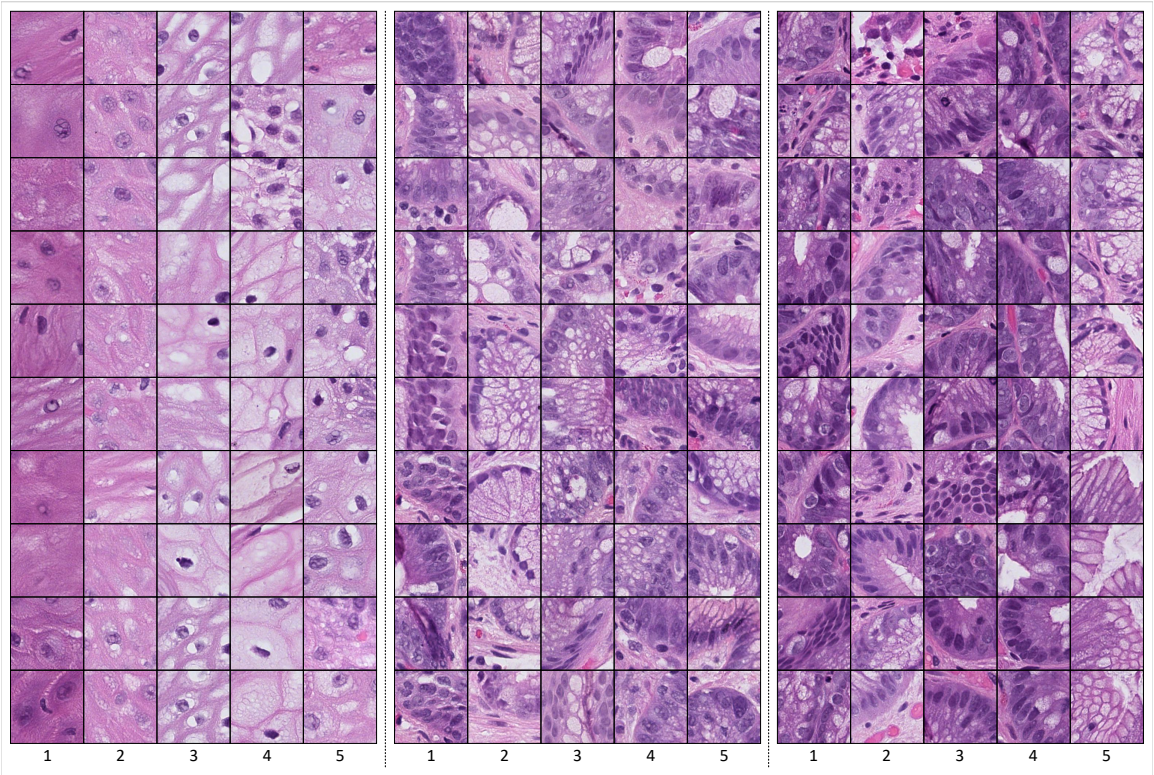

Figure 4.4: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the BoVW-CAE

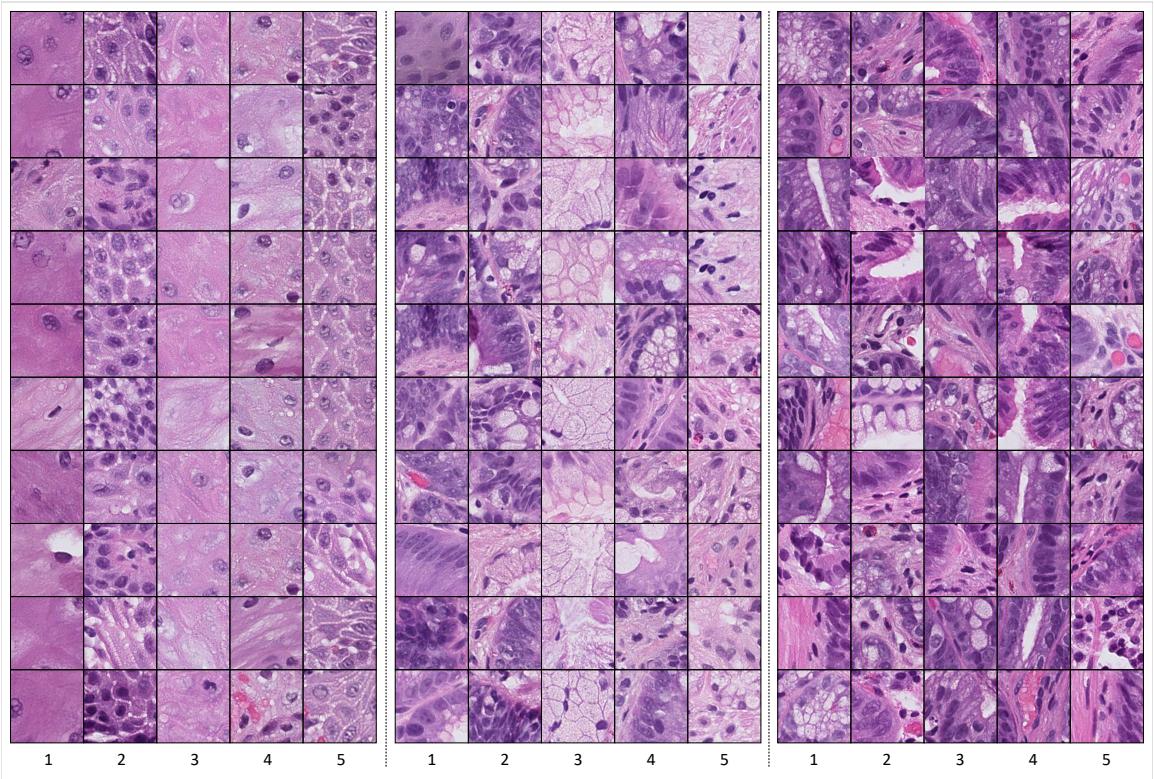

Figure 4.5: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the BoVW-BiGAN

4.4 Discussion

In this chapter, we investigated the BoVW capability for representation learning of histology images to classify dysplastic and non-dysplastic BE. We used a two-step process, in which, in the first step, local image features are quantified and then clustered to shape codewords. Finally, each image is represented as an order-less histogram of codeword frequencies. A decision fusion model was trained on image-level histograms in the second step to output the final labels of new WSIs. The results demonstrated the ability of unsupervised feature extraction from WSIs if an appropriate setting is chosen. This is an important contribution because we provided an informative representation of WSIs, employing an unsupervised framework that avoids manually image annotation.

Since the number of clusters is a critical parameter in some clustering algorithms such as K-means, we evaluated different values to select a decent number of visual words that generate better classification results. As shown, both feature learning approaches, CAE and BiGAN, generated very similar results given a specific number of clusters (here 100). However, applying BoVW on deep local features learned by CAE outperforms the same model on BiGAN-derived features considering different values for the number of clusters.

Despite producing satisfactory results, the BoVW approach does not consider the spatial arrangement of visual words, which leads to some information loss. In some cases, the contribution of visual word distribution to the outcome is negligible and hence can be ignored, while in some cases, such information loss leads to a significant deterioration of model performance.

5 | Neural Image Compression

5.1 Background

Although WSIs are composed of millions of pixels, only a small portion of these pixels tend to be meaningful regarding relevant image-level metadata (e.g., the image label in the classification problem or time in survival analysis) [45]. In the current published literature, analysis of WSIs has been addressed by different machine learning paradigms. In the fully-supervised paradigm, [24,25,26], pixel-wise annotation of images and extracting image patches from annotated regions for further analysis is an approach to nullify the dimensionality curse of WSIs. Employing the fully-supervised approach requires accepting some simplifying assumptions. One of the most common assumptions states that the signal associated with the image-level metadata has a patch-level representation and can be fully recognized at a low level of abstraction [45]. This assumption creates the need for patch-level annotation, which means engaging in a labor-intensive, time-consuming, and error-prone process. However, it is obvious that this method is applicable only in cases where the relationship between image patterns and image-level metadata is known. In such cases, the spatial relationship between patches is lost in the process of patch-level annotation. If the relationship between image patterns and image-level metadata is not fully understood, as may be the case with pathology slides for newly discovered diseases, preserving the spatial arrangement of local image features may aid in discovering and understanding new disease processes.

Another assumption states that although the image-level metadata signals exist at a low level of abstraction, a human annotator cannot recognize them. According to this assumption, patches' mere presence is sufficient to ensure the existence of true signal evidence associated with the image-level metadata even if information regarding the spatial arrangement of patches is lost. Based upon this assumption, the gigapixel images are assumed to be a set of bags, each containing many patches that only some of them trigger the true signal. This scenario is called Multiple Instance Learning (MIL) [28, 61]. One of the goals in MIL is detecting critical patches. This approach cannot provide any analysis

beyond the patch-level due to not considering the spatial relationship between patches. However, methods such as Attention-Based MIL [47] have been proposed to improve the interpretability of MIL.

Tellez et al. [45] proposed Neural Image compression (NIC) for gigapixel image analysis. NIC sidesteps the steep computational requirements needed to train neural networks with whole slide images by organizing the patch-level representation vectors of each WSI, given their spatial arrangement. This approach creates highly compact representations that are more amenable to train neural networks. This technique creates highly compact representations of gigapixel images that are more amenable to training neural networks. Inspired by this method and utilizing the concept of visual words, we proposed Cluster-Based Neural Image Compression (CBNIC), which leverages the advantages of NIC and also incorporates the information regarding codeblocks into the model. The performance of the proposed approach is evaluated for the diagnosis of dysplastic and non-dysplastic BE on WSIs.

5.2 Model Development

In this model, a gigapixel image $w \in \mathbb{R}^{R \times C \times 3}$ (R : number of rows, C : number of columns, and three color channel (RGB)) which its feeding into a CNN carries a steep computational cost, if not infeasible given current technology, is compressed in such a way that a CNN can be trained directly on these images in a lower computational cost. CBNIC consists of two main phase: encoding of the images and training a CNN on compressed images. Figure 5.1 represents the overview of the model. Different steps of this model are explained in more detail as follows.

5.2.1 Image Compression

Codebook Learning

In this step, the local image descriptors of images are identified. After dividing the gigapixel image w into a set of high-resolution tissue tiles $x_{ij} \in \mathbb{R}^{P \times P \times 3}$ sampled from the i^{th} row and j^{th} column of an uniform grid of square patches of size P using a stride of S throughout w, each tissue tile x is mapped into a low-dimensional embedding vector size E independently. This study employs CAE and BiGAN trained in an unsupervised fashion to map each high-resolution tissue tile into a low-dimensional embedding space. Then, a k-means clustering algorithm is employed to cluster extracted features into several clusters (see Fig 5.1, A-1). Each cluster is indeed a visual descriptor called a visual word or codeblock, which are components of a codebook. Selection of the number of clusters (codebook size) is an

important decision in codebook construction. This parameter should be guessed/optimized and then imported to the model as an an input.

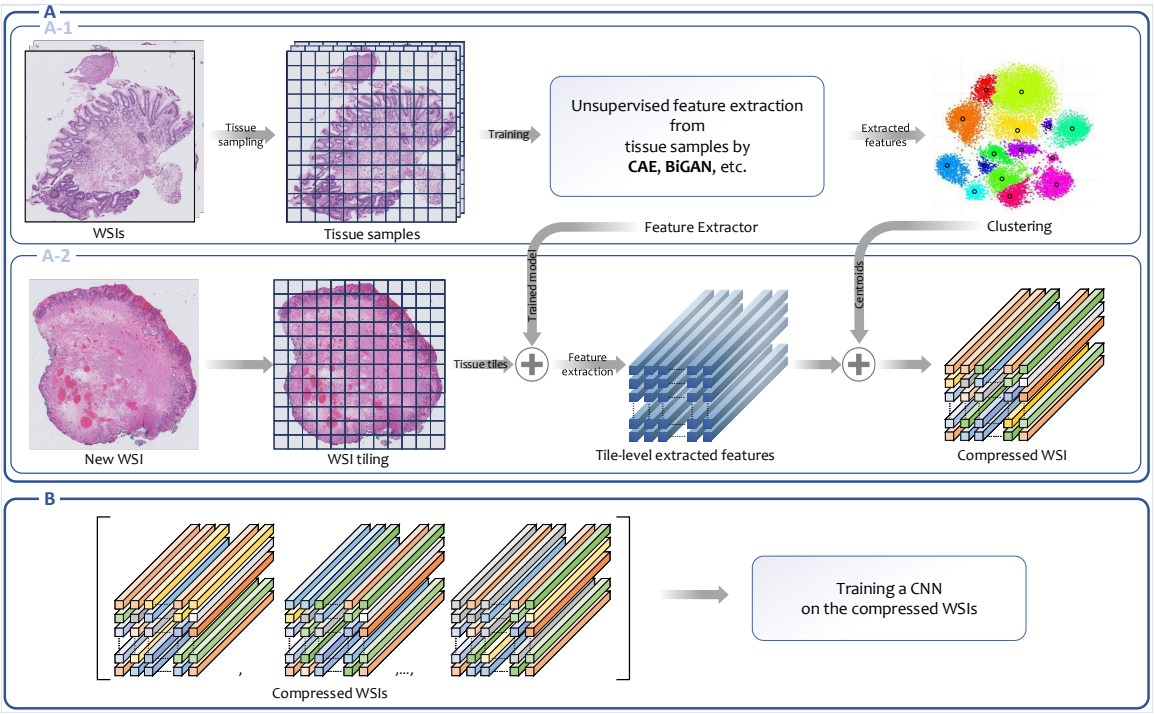

Figure 5.1: Overview of CBNIC approach

WSI Encoding

After codebook learning, each tissue tile of a WSI gets associated with a visual word. Then, its embedding vector is replaced by the representation vector of the associated visual word. Finally, the visual words are organized following the same spatial arrangement of corresponding tissue tiles as in the WSI w (see Fig 5.1, A-2). In this way, the information regarding codeblocks is incorporated into the model. As a result of this phase, gigapixel image $w \in \mathbb{R}^{R \times C \times 3}$ is compressed as $w' \in \mathbb{R}^{\frac{R}{S} \times \frac{C}{S} \times E}$.

5.2.2 Training of a CNN on the compressed WSIs

Now that WSIs have been compressed, a CNN can be trained on the compressed images to predict the output for a new WSI. A CNN can detect both local and global discriminative visual features from the images and NIC preserving the spatial arrangement of tissue patches guarantees that the CNN is provided with both local and global information from gigapixel images.

5.3 Experiments and Results

5.3.1 Architecture of CNN

A shallow CNN composed of six 2D convolutional layers was employed for classification of the compressed WSIs. We used kernel size 3, pooling 1, and stride 1 to keep the feature resolution unchanged. Convolution layers are followed by a softmax layer that outputs the class probabilities. The rectified linear unit (ReLU) [14] is employed as the activation function. Also Batch Normalization [62] was applied after ReLU of every trainable layer. We utilized Global Attention Pooling to highlight the informative patches for the classification of esophageal WSIs. Global Attention Pooling computes the contribution of each visual word by learning the weights for the corresponding features vector and pools them from all the visual words present in the given image to provide more optimal representation for an image-level task. The importance of each visual word is estimated based upon its embedding vector and its neighboring visual words.

5.3.2 Experimental Results

The performance of a classification model is highly correlated with the degree of separability between different classes. Before applying classification algorithms on encoded WSIs, we visualized image-level representations provided by CBNIC using PCA method to understand better how well each method characterizes the visual content of histopathology images (see Figure 5.2). What can be deduced from the graphs is that WSIs in squamous and dysplastic BE have been encoded relatively separately from each other, while non-dysplastic BE images have a sort of confusion with both these classes.

Classification results can further refine our findings from the PCA plots. To evaluate the model quantitatively, five standard metrics were used for classification under a 1-vs-rest strategy: accuracy, precision, recall, specificity, and F1 score. To estimate 95% CIs, bootstrapping was used for all metrics.

After applying the sliding window method on 650 WSIs from 120 unique patients, a total of 2135 big patches (5000×5000 pixels) were generated, of which 793 (37.1%) were in the squamous class, 606 (28.4%) were in Barrett's class, and 736 (34.5%) were in the dysplastic BE class. Of the independent testing set of 321 images, 142 (44.2%) squamous, 74 (23.1%) Barrett's, and 105 (32.7%) dysplastic BE images were used to evaluate trained models and to analyze the classification performance from both quantitative and qualitative aspects. Image augmentation was also performed by horizontal flipping, random 90-degree rotations, and image mirroring during training to prevent the CNN from over-fitting.

The classification results of esophageal WSIs in squamous, dysplastic BE, and non-

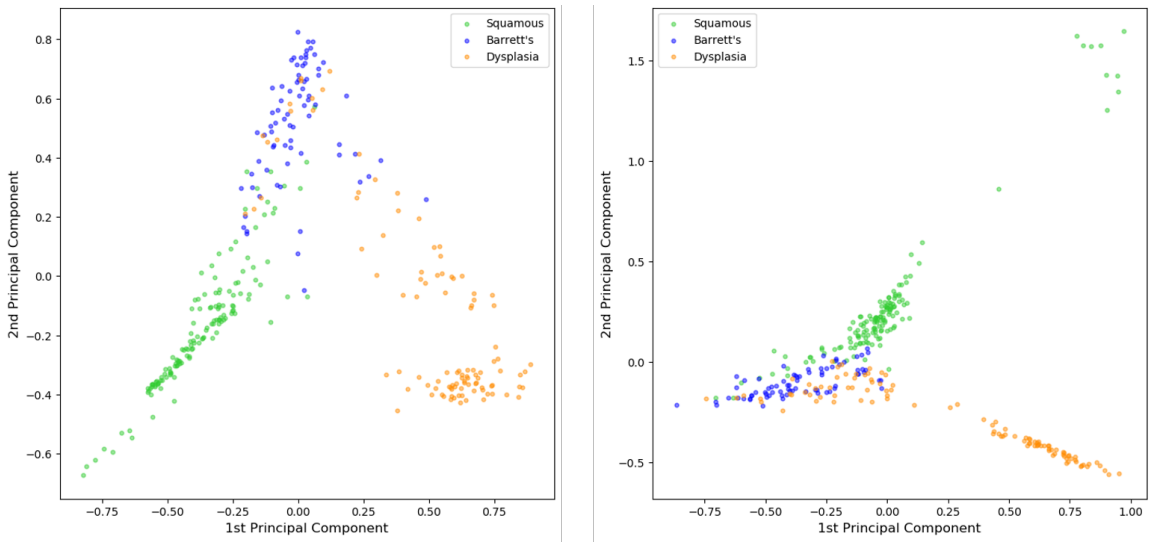

Figure 5.2: PCA plot for WSIs encoded using (left) CBNIC-CAE, (right) CBNIC-BiGAN.

dysplastic BE for different models are summarized in Table 5.1. The reported values are averages with 95% confidence intervals. For computing confidence intervals, numbers greater than one were truncated to 1.

As shown in Table 5.1, the CBNIC model on CAE-derived image features outperforms the model on BiGAN-based features given the weighted average of F1 score as an evaluation metric. The weighted average of F1 score for the model trained on encoded WSIs using CAE is 0.889 (95% CI, 0.884-0.894) vs. 0.858 (95% CI, 0.855-0.862) for the model trained on encoded WSIs using BiGAN. Tables 5.2 and 5.3 shows the confusion matrix of BoVW-CAE and BoVW-BiGAN respectively. As shown in both models, the dysplastic BE has mostly confusion with non-dysplastic BE.

Table 5.1: Results of WSI classification using CBNIC approach

Class	Metric	Model	
		CBNIC-CAE	**CBNIC-BiGAN**
Squamous	Accuracy	0.936 (0.931, 0.941)	0.930 (0.927, 0.933)
	Precision	0.939 (0.930, 0.948)	0.926 (0.922, 0.931)
	Recall	0.918 (0.912, 0.925)	0.914 (0.909, 0.918)
	Specificity	0.950 (0.943, 0.958)	0.942 (0.939, 0.946)
	F1 score	0.928 (0.922, 0.933)	0.920 (0.916, 0.923)
Barrett's	Accuracy	0.898 (0.893, 0.902)	0.895 (0.892, 0.898)
	Precision	0.733 (0.721, 0.745)	0.714 (0.705, 0.722)
	Recall	0.884 (0.868, 0.900)	0.907 (0.901, 0.914)
	Specificity	0.902 (0.896, 0.908)	0.892 (0.888, 0.895)
	F1 score	0.798 (0.788, 0.807)	0.798 (0.792, 0.804)
Dysplasia	Accuracy	0.939 (0.936, 0.942)	0.890 (0.887, 0.894)
	Precision	0.964 (0.959, 0.968)	0.901 (0.895, 0.907)
	Recall	0.845 (0.836, 0.853)	0.746 (0.738, 0.754)
	Specificity	0.985 (0.983, 0.987)	0.960 (0.957, 0.963)
	F1 score	0.899 (0.894, 0.905)	0.816 (0.810, 0.822)
Weighted Average	Accuracy	0.928 (0.925, 0.932)	0.909 (0.906, 0.911)
	Precision	0.900 (0.895, 0.905)	0.870 (0.867, 0.873)
	Recall	0.887 (0.881, 0.892)	0.857 (0.854, 0.861)
	Specificity	0.950 (0.947, 0.954)	0.936 (0.935, 0.938)
	F1 score	0.889 (0.884, 0.894)	0.858 (0.855, 0.862)

Table 5.2: Confusion matrix of CBNIC-CAE

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	139 (0.979)	3 (0.021)	0 (0.000)
	Non-dysplastic BE	3 (0.041)	69 (0.932)	2 (0.027)
	Dysplastic BE	1 (0.010)	15 (0.143)	89 (0.848)

Table 5.3: Confusion matrix of CBNIC-BiGAN

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	132 (0.930)	4 (0.028)	6 (0.042)
	Non-dysplastic BE	4 (0.054)	67 (0.905)	3 (0.041)
	Dysplastic BE	7 (0.067)	21 (0.200)	77 (0.733)

As the number of clusters is an important parameter in the performance of clustering methods, we evaluated different numbers of clusters (codeblocks) for both approaches to pick a decent number of codeblocks. As shown in Figure 5.3, the CBNIC-CAE outperforms the CBNIC-BiGAN regardless of the number of clusters. Furthermore, by changing the number of clusters, no significant change observed in the the CBNIC-CAE performance. In contrast, in the CBNIC-BiGAN, after a significant increase in the model performance due to increasing the number of clusters from 50 to 100, the value of the weighted F1 score decreases as a result of increasing the number of clusters and again with further increase in the number of clusters, the model performance is improved. Therefore, to have a less

complex model, 100 is determined as an optimal value for the number of clusters, although 250 and 300 clusters lead to very similar classification results. Also, for CBNIC-CAE, 100 is selected as the optimal number of clusters.

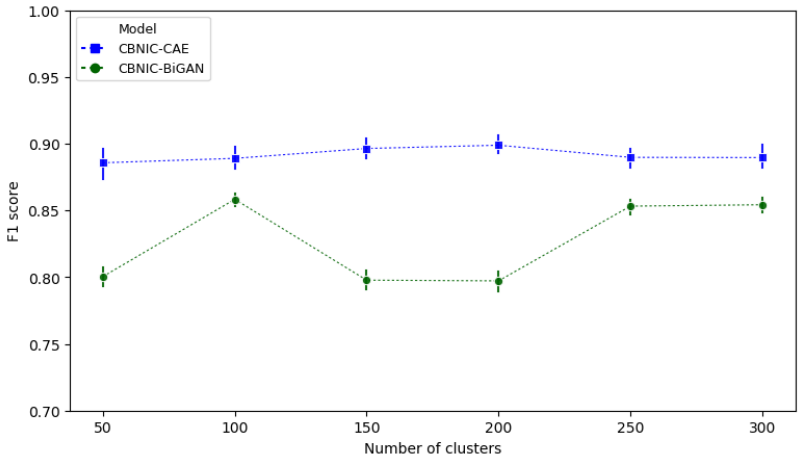

Figure 5.3: The effect of the number of clusters on the performance of CBNIC

Figure 5.4 shows the performance of the CBNIC model vs. BoVW. As indicated, employing the CBNIC for the classification of esophageal WSIs has generated better results considering both CAE and BiGAN. Given the inherent differences between the two approaches, explaining the origin of this superiority is out of this dissertation's scope. One of the advantages of CBNIC over BoVW is taking into account the spatial arrangement of image patches. We investigated this characteristic's contribution in the final result by running the CBNIC on images lacking this feature. To do so, after extracting the embedding vectors from tissue patches and learning the visual codebook, the codeblocks' representation vectors were organized in random spatial locations to generate the compressed images. Figure 5.5 shows the comparison between CBNIC model on images compressed with and without preserving the spatial arrangement between the image patches. As can be seen, the CBNIC model on images compressed preserving the spatial arrangement of patches, outperforms the other model. This experiment only shows the positive impact of spatial arrangement of patches on the final result. But cannot provide any conclusion regarding quantity of this impact because the spatial location of patch-level embedding vectors were randomly shuffled and we had no control over the degree of disarray. The difference in the degree of deterioration of the results for different values of the number of clusters can be explained in this way.

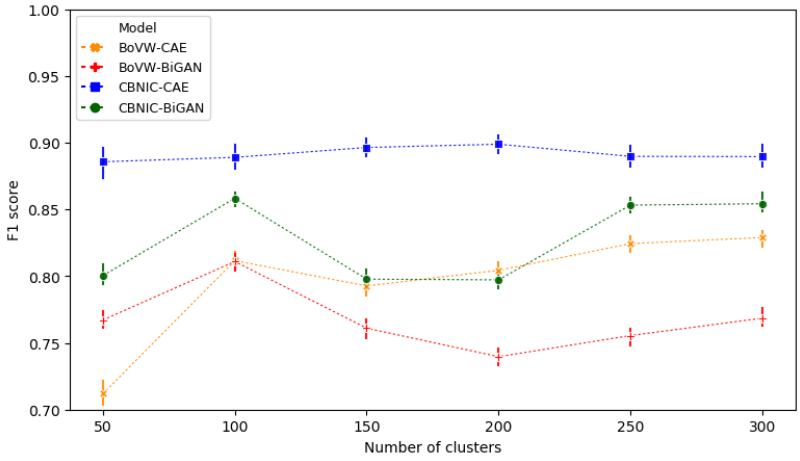

Figure 5.4: Comparison of the performance of BoVW and CBNIC

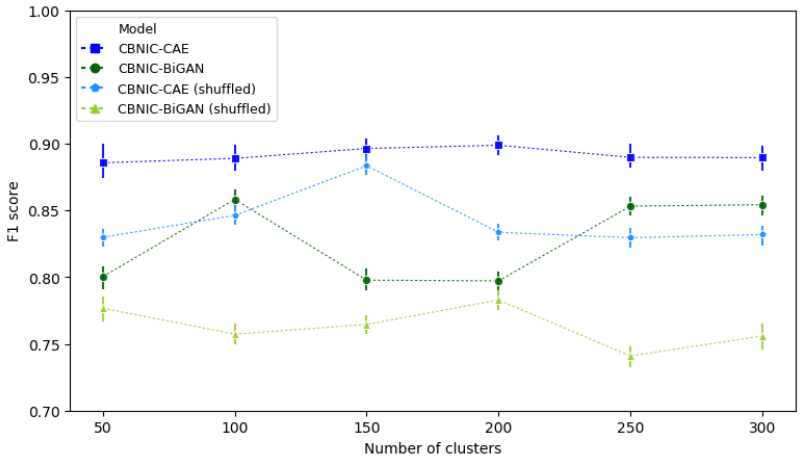

Figure 5.5: The contribution of spatial arrangement of tissue patches in the performance of CBNIC

5.3.3 CBNIC vs NIC

As explained earlier, the contribution of CBNIC over NIC is incorporating the concept of codeblock into the model instead of compressing the WSIs using patch-level embedding vectors directly. In this section, the contribution of this change is evaluated. Tables 5.4 and 5.5 show the confusion matrices of NIC given the image features learned by CAE, and BiGAN respectively.

Table 5.4: Confusion matrix of NIC-CAE

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	112 (0.789)	15 (0.106)	15 (0.106)
	Non-dysplastic BE	14 (0.189)	60 (0.811)	1 (0.000)
	Dysplastic BE	24 (0.229)	19 (0.181)	62 (0.590)

Table 5.5: Confusion matrix of NIC-BiGAN

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	104 (0.732)	33 (0.232)	5 (0.035)
	Non-dysplastic BE	3 (0.041)	70 (0.946)	1 (0.014)
	Dysplastic BE	22 (0.210)	40 (0.381)	43 (0.410)

Comparing these results with the results of the CBNIC model (Tables 5.2 and 5.3), it can be deduced that the CBNIC model outperforms the NIC model for the classification of esophageal WSIs. We employed bootstrapping approach (100 iterations) to test the significance of the results. Figure 5.6, represents the boxplot of F1 score values on different iterations for both NIC and CBNIC.

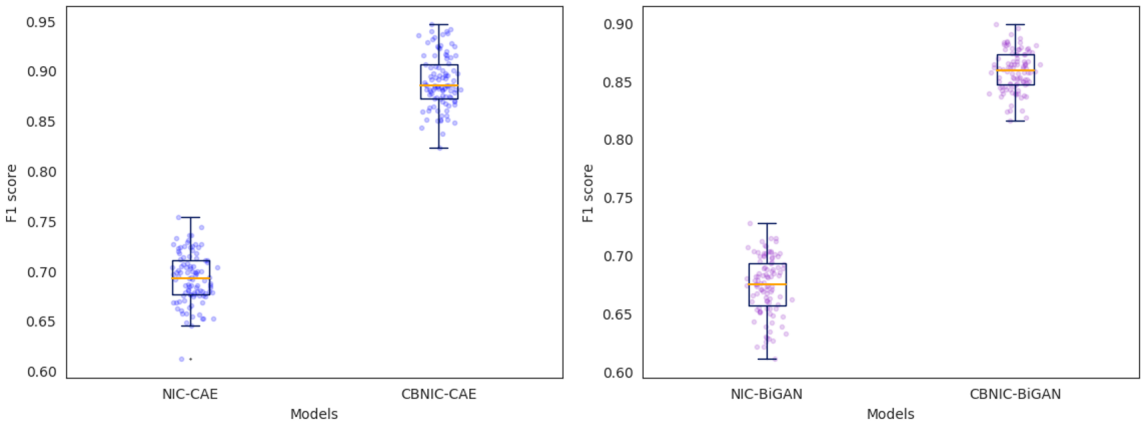

Figure 5.6: Comparison of the performance of NIC and CBNIC for classification of WSIs encoded by (left) CAE, (right) BiGAN

As shown, the performance of CBNIC is significantly better than NIC, and this is evidence for the contribution of incorporating codeblock concept into the NIC.

5.3.4 Codeblock Analysis

After training the classifier on compressed images, a weight is learned for each codeblock that is an estimate of how important it is in the image-level inference. By aggregating the weights from all images in the test set for every codeblock and averaging, a value is obtained indicating the relative importance of given codeblock. Figures 5.7 and 5.8 show some randomly selected image patches associated with codeblocks of the highest importance for CAE-derived and BiGAN-derived features, respectively.

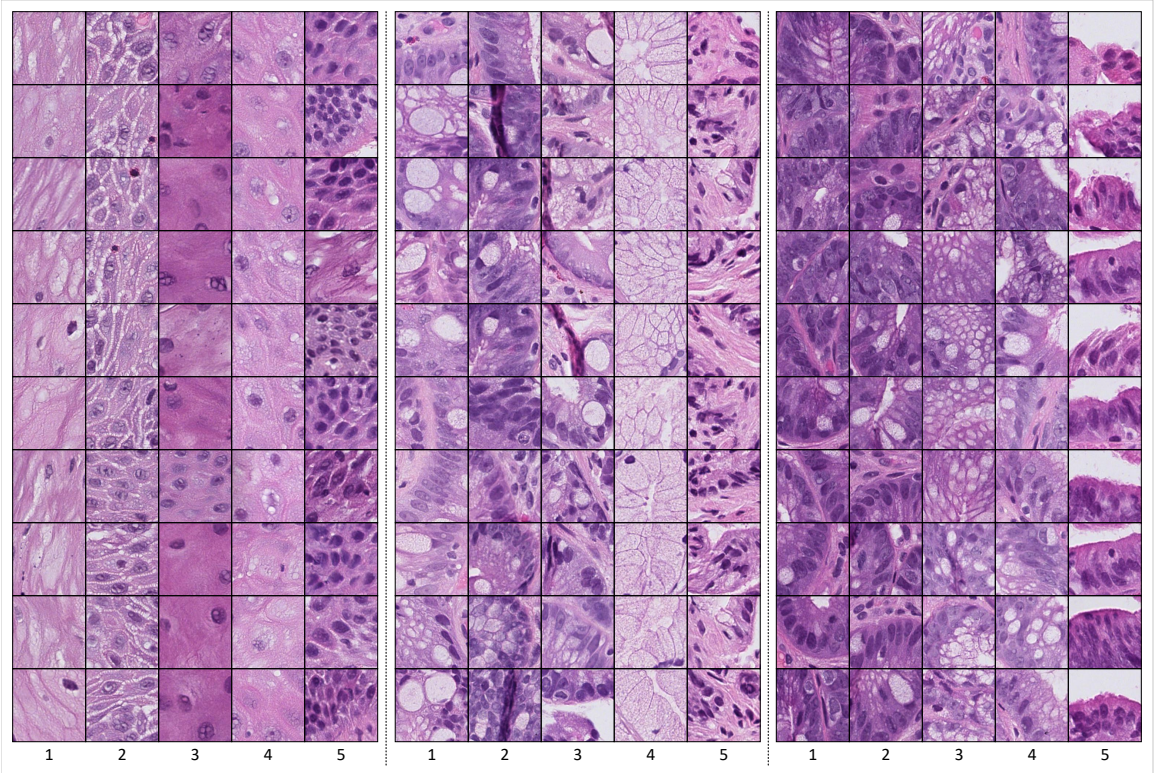

Figure 5.7: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the CBNIC-CAE

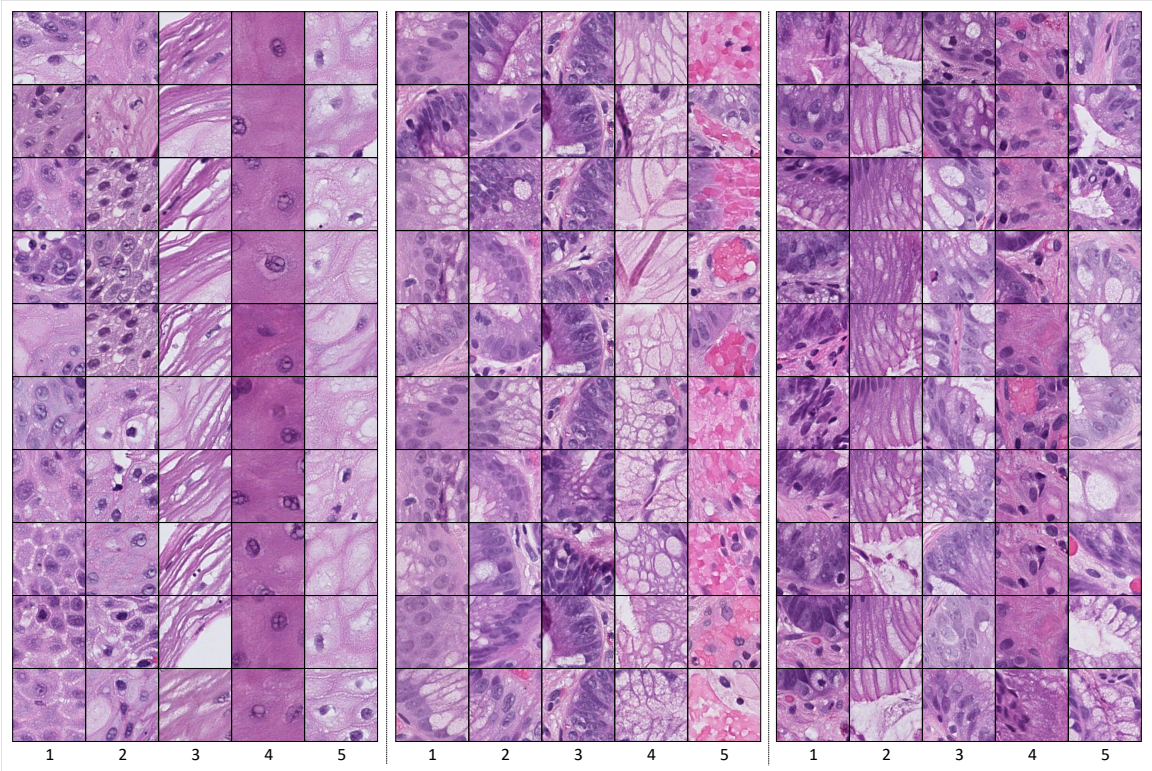

Figure 5.8: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the CBNIC-BiGAN

A medical expert reviewed these sampled tissue tiles. In both models, tissue tiles for squamous were indicative of normal esophageal histology showing squamous cells that are flat, thin cells lining the esophagus's surface. For non-dysplastic BE, tiles showed goblet cells and columnar epithelium that are normally only present in the small intestine, but with Barrett's, the normal esophageal lining shows the presence of these mucin-containing goblet cells and columnar epithelial cells. For dysplastic BE, crowded hyperchromatic nuclei were visualized that are known to be present among pre-cancerous dysplastic cells.

5.4 Discussion

This chapter proposed CBNIC as a cluster-based version of the NIC for representation learning of high-resolution histopathology images. The NIC approach compresses gigapixel images by learning the embedding vector of image patches in an unsupervised fashion and organizing them following the same spatial arrangement as the original image. While, in the CBNIC, a codebook of visual words is learned, and the embedding vectors of image patches are replaced by the embedding vector of the corresponding visual word, and then the gigapixel images are compressed the same way as the NIC. The performance of the learned representations by CBNIC was evaluated through an end-to-end supervised framework. The

results demonstrated the ability of the CBNIC approach to detect dysplastic and non-dysplastic BE on WSIs. The results also show that the CBNIC significantly outperforms the NIC.

The CBNIC and BoVW (discussed in Chapter 4) are somehow similar in terms of utilizing the visual words for representation learning of images, however, from different angles. While the BoVW considers frequencies of visual words, the CBNIC uses the visual words' representation vectors to encode the images. If we accept this similarity, we can say that CBNIC can be a solution to relive the lack of considering the spatial information of local features in the BoVW. As illustrated, the CBNIC outperforms the BoVW, although demonstrating that the reason for this superiority is considering the spatial information of local image features in the CBNIC compared to the BoVW is not within the scope of this study owing to the differences in the nature of the two models. However, we showed the positive contribution of considering the spatial information of image patches by running the CBNIC on the compressed images, which in the image patches are randomly shuffled and no longer are in their original spatial position. As illustrated, the classification results of the compressed images preserving the spatial information of local image features is significantly better than the results of the classification of images compressed without preserving the spatial information.

As demonstrated, the CBNIC model on the images compressed using CAE-based image features has a stable prediction ability for the different number of clusters. In contrast, its performance on BiGAN-derived image features varies depending on the number of clusters. Furthermore, the model on CAE image features significantly outperforms the same model on the BiGAN-based ones. Given that the k-means performance is highly dependent on its initialization, one might think that the superiority of the model on the CAE-based features over the BiGAN-based features stems from the proper initialization of k-means clustering on CAE features. Although this argument can be logically valid, here is less likely to be the case because the results show the superiority of the CBNIC on CAE-derived image features for all experimented cluster numbers. Besides, an attempt has been made to choose a relatively appropriate initialization point for each clustering run using the trial and error method. This further calls into question the validity of the raised argument.

6 | Graph Neural Networks

6.1 Background

As said earlier, CNNs have gained increasing attention in image classification due to their ability to capture image feature representations. However, the gigapixel high-resolution images cannot be employed directly to train a CNN due to high computational cost. Some approaches such as BoVW [36], MIL, etc., have been employed to tackle this problem, but they lack the spatial relationships between nearby image patches. Hypothetically, the closer the two image patches are spatially, the more similar they are, concerning the image-level metadata (e.g., image label or survival time). Graph convolutional networks (GCNs) [63] are an efficient architecture that can effectively capture such spatial proximity information by modeling relations between image patches using vertexes. Therefore, GCNs can be utilized to model spatial arrangement between image patches extracted from the WSIs, which fail to be considered in CNNs.

In the literature, very little attention has been devoted to the application of GCNs to analyze whole-slide histopathology images. Based on our best knowledge, a few studies have been accomplished to explore this possibility. Sureka et al. [64] used GCNs for the classification of histopathology images. They encoded histology tissue as a graph of nuclei. They used an attention-based architecture to provide an interpretable map highlighting each nucleus's contribution and neighborhood in the final diagnosis. Zhu et al. [65] benefited from GCNs for grading of Colorectal cancer. They used a GCN to convert each large histology image into a graph, where nuclei are nodes of a graph, and cellular interactions are denoted as edges between them according to node similarity. Adnan et al. [40] employed GCNs to learn a representation of lung cancer WSIs. They sampled relevant patches and utilized graph neural networks to capture relations among sampled patches to aggregate the WSI information into a single vector representation. They evaluated the quality of learned representations by classification of WSIs in an end-to-end framework. Konda et al. [66] also showed the ability of GCNs in the classification of histopathology images of colon cancer and breast cancer.

This chapter proposes a novel architecture based upon graph neural networks to learn a representation of WSIs. In this model, each WSI is encoded as a graph in which the nodes are visual words connected using vertexes to capture their spatial proximity information.

6.2 Review of GCNs

6.2.1 Graph Representation

A graph is defined as an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ represents the set of vertices consisting of nodes $\{v_1, ..., v_n\}$, and $\mathcal{E}$ denotes to the set of edges. Each node v_i has a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and the entire feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ includes the feature vectors from all nodes $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$. In such graph, $W \in \mathbb{R}^{n \times n}$ is a weighted adjacency matrix where w_{ij} denotes the edge weight between nodes v_i and v_j.

In our context, as shown in Figure 6.1, each WSI is modeled as a complete graph in which each image patch is treated as a vertice, and a unique edge connects every pair of distinct vertices. The weight of each edge w_{ij} is a function of similarity between corresponding nodes' feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ and also their spatial distance (see Equation 6.1).

$$w_{ij} = \frac{sim(\mathbf{x_i}, \mathbf{x_j})}{1 + dist(v_i, v_j)} \tag{6.1}$$

Where $sim(\mathbf{x_i}, \mathbf{x_j})$ is similarity between feature vectors of nodes v_i and v_j and $dist(v_i, v_j)$ is spatial distance between these two nodes.

Choice of the proper distance metric and calculation of similarity in high dimensional applications, which is the case here, is very heuristic [67]. To calculate the similarity between $\mathbf{x_i}$ and $\mathbf{x_j}$, we considered the average of cosine similarity and similarity calculated based upon $L^2 - norm$ distance to benefit from advantages of both measures (see Equation 6.2).

$$sim(\mathbf{x_i}, \mathbf{x_j}) = \frac{\frac{\mathbf{x_i}.\mathbf{x_j}}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|} + \frac{1}{1+\|\mathbf{x_i}-\mathbf{x_j}\|}}{2} \tag{6.2}$$

The Euclidean distance between two nodes v_i and v_j is derived from Equation 6.3.

$$dist(v_i, v_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{6.3}$$

Where x_i and y_i are coordinates of the center of image patch v_i.

6.2.2 Graph Convolution

In this study, two types of graph convolution are employed as follows:

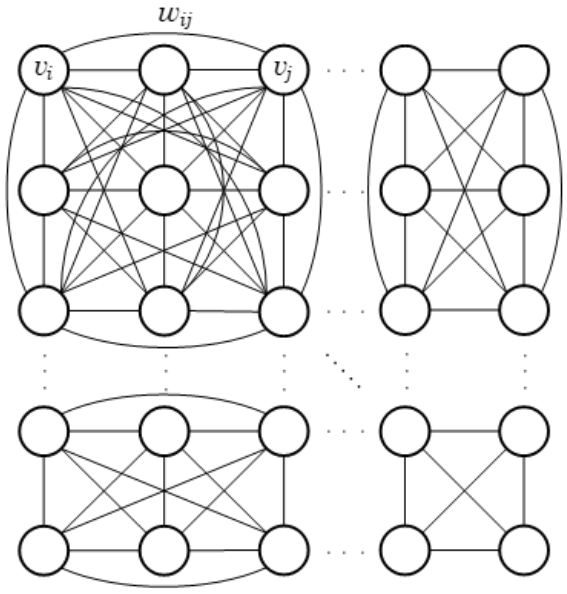

Figure 6.1: Graph representation

Spatial Convolution

Generally speaking, the spatial Convolution is propagation of node features to neighborhood nodes followed by activation function in a message passing network [68]. A message passing graph neural network can be described as follows:

$$h_i{}^{t+1} = U_t(h_i{}^t, \Theta_{j \in N_i}(M_t(h_i{}^t, h_j{}^t, e_{ij}^t, w_{ij}^t))) \tag{6.4}$$

Where $h_i{}^t \in \mathbb{R}^F$ denotes features vector of node i in layer t and $h_i^0 = \mathbf{x}_i$, N_i denotes set of neighbors of node i in graph $\mathcal{G}$, $e_{ij}^t \in \mathbb{R}^D$ is features vector of the edge between node i and node j in layer t, and $w_{ij}^t \in \mathbb{R}$ denotes the weight of this edge in the same layer. Θ denotes a differentiable, permutation invariant function, e.g., sum, mean or max. The message function M_t and vertex update functions U_t are learned differentiable functions [68].

Spectral Convolution

In spectral domain, eigendecomposition of graph Laplacian is employed to filter the signal on the graphs. A graph Laplacian is defined as $L = D - W$ or in the normalized form it is $L = I_n - D^{-1/2}WD^{-1/2}$ where $W \in \mathbb{R}^{n \times n}$ denotes the adjacency matrix of graph $\mathcal{G}$, I_n denotes the identity matrix and $D \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix where each entry on the diagonal is equal to the row-sum of the adjacency matrix: $d_i = \sum_j w_{ij}$. Since L is positive semidefinite, it can be decomposed into $L = U\Lambda U^\top$, where $U = [u_1, ..., u_n] \in \mathbb{R}^{n \times n}$

is the eigenvectors matrix and $u_i \in \mathbb{R}^n$ are graph Fourier modes. Also, $\Lambda = diag(\lambda) \in \mathbb{R}^{n \times n}$ where $\lambda = [\lambda_1, ..., \lambda_n]$ and λ_i is a real non-negative eigenvalue associated with u_i [69].

The convolution operator in graph $\circledast$ is defined in the frequency domain as follows [70]:

$$x \circledast y = U(U^\top x \odot U^\top y) \tag{6.5}$$

Where $\odot$ represents the element-wise Hadamard product. Also, the graph Fourier transform of signals x and y are defined as $U^\top x$ and $U^\top y$ respectively. As shown in Equation 6.6, the spectral convolution on graphs in frequency domain is defined as the multiplication of any unidimensional signal x on graph by a filter g_θ [70].

$$g_\theta \circledast x = g_\theta(L)x = g_\theta(U\Lambda U^\top)x = U g_\theta(\Lambda)U^\top x \tag{6.6}$$

Where $g_\theta(\Lambda) = diag(\theta) = diag(\mathcal{F}(\lambda))$ where $\mathcal{F}$ is a desired filter function [69].

Evaluating the Equation 6.6 is computationally expensive as calculation of Fourier and inverse Fourier transform by matrix multiplication of U and $U^\top$ is in $\mathcal{O}(n^2)$ [63]. Parameterizing the $g_\theta(L)$ as a polynomial function that can be computed recursively from L is a solution. To get around this problem, Hammond et. al [71] suggested that $g_\theta(\Lambda)$ can be approximated by a truncated expansion of Chebyshev polynomials $T_k(x)$ [63]. The Chebyshev polynomial is obtained from the following recurrence relation:

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \qquad T_0(x) = 1\ , \quad T_1(x) = x \tag{6.7}$$

According to 6.7, approximation of $g_\theta(\Lambda)$ is as follows [63]:

$$g_{\theta'}(\Lambda) \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\Lambda}) \tag{6.8}$$

Where $\tilde{\Lambda} = \frac{2}{\lambda_{max}}\Lambda - I_n$. Now, the Equation 6.6 can be rewritten as Equation 6.9.

$$g_{\theta'} \circledast x = U g_{\theta'}(\Lambda)U^\top x \approx \sum_{k=0}^{K} \theta'_k T_k(U\tilde{\Lambda}U^\top)x = \sum_{k=0}^{K} \theta'_k T_k(\tilde{L})x \tag{6.9}$$

6.2.3 Graph Pooling

Graph pooling refers to operations applied to reduce the number of nodes or downsample the node features in a graph and have a similar role to the pooling in traditional CNNs. Since pooling computes a coarser version of the graph at each step, ultimately resulting in a single vector representation, it is usually applied to graph-level inference problems

such as graph classification. Different types of graph polling layers such as Max Pooling, Mean pooling, Sum pooling, and Global Attention Pooling can be employed to pool the node feature vectors in a single representation vector. In this study, we utilized Global Attention Pooling to highlight the informative patches for the classification of esophageal WSIs. Global Attention Pooling computes each node's contribution by learning the weights for the corresponding features vector and pools them from all the nodes to provide a more optimal representation for a graph-level task.

6.3 Model Development

We applied CBGCN for the classification of WSIs. In this model, each WSI is encoded as a graph as explained earlier, and then a GCN is trained on WSI-level graphs to predict the label of new WSIs. Figure 6.2 illustrates an overview of the proposed approach. Also, different steps of this model are explained in more detail as follows:

6.3.1 Image Compression

Codebook Learning

In this step, the local image descriptors of images are identified. After dividing the gigapixel image w into a set of high-resolution tissue tiles $x_{ij} \in \mathbb{R}^{P \times P \times 3}$ sampled from the i^{th} row and j^{th} column of an uniform grid of square patches of size P using a stride of S throughout w, each tissue tile x is mapped into a low-dimensional embedding vector size E independently. This study employs CAE and BiGAN trained in an unsupervised fashion to map each high-resolution tissue tile into a low-dimensional embedding space. Then, a k-means clustering algorithm is employed to cluster extracted features into several clusters (see Fig 6.2, A-1). Each cluster is indeed a visual descriptor called a visual word or codeblock, which are components of a codebook. Selection of the number of clusters (codebook size) is an important decision in codebook construction. This parameter should be guessed/optimized and then imported to the model as an an input.

WSI Encoding

After codebook learning, the embedding vector of each tissue tile $\mathbf{x}$ gets associated with a visual word. Then, the embedding vectors of visual words are considered feature vectors of nodes in a complete graph (see Fig 6.2, A-2). In this way, the information regarding codeblocks is incorporated into the model.

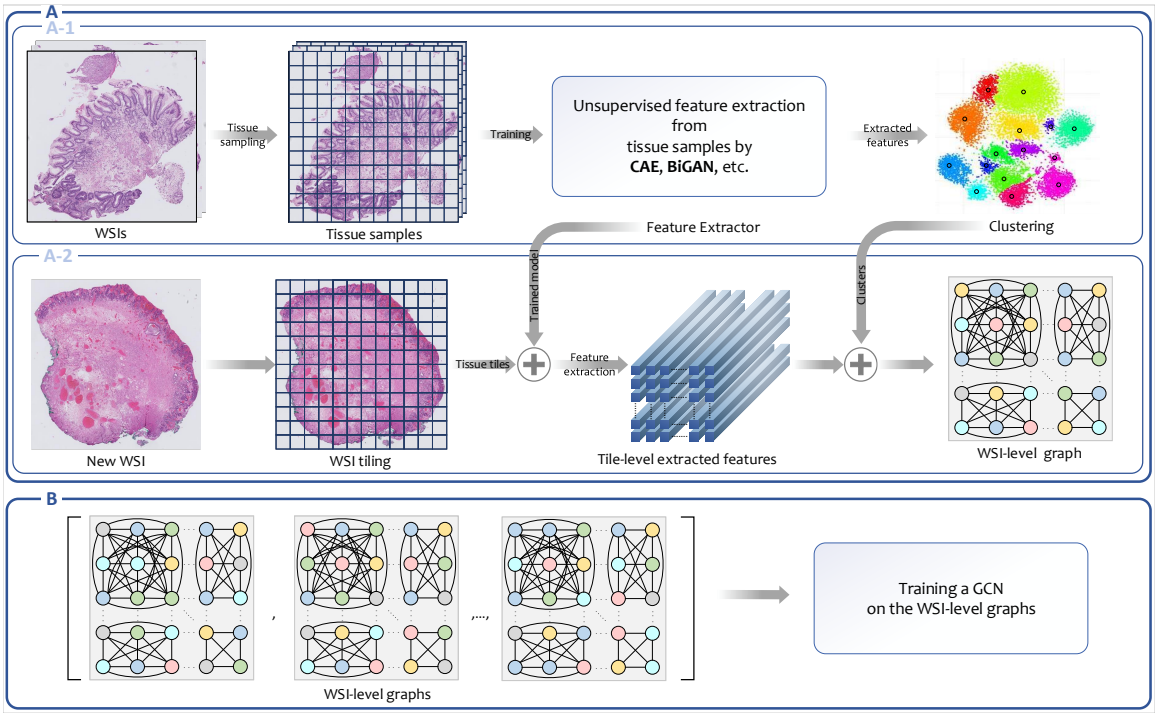

Figure 6.2: Overview of CBGCN approach

6.4 Experiments and Results

6.4.1 Architecture of GCN

The GCNs employed in this study were composed of four graph convolutional layers followed by pooling and softmax layer.

6.4.2 Experimental Results

After applying the sliding window method on 650 WSIs from 120 unique patients, a total of 2135 big patches (5000 × 5000 pixels) were generated, of which 793 (37.1%) were in the squamous class, 606 (28.4%) were in non-dysplastic BE class, and 736 (34.5%) were in the dysplastic BE class. Of the independent testing set of 321 images, 142 (44.2%) squamous, 74 (23.1%) Barrett's, and 105 (32.7%) dysplastic BE images were used to evaluate trained models and to analyze the classification performance from both quantitative and qualitative aspects. Image augmentation was also performed by horizontal flipping, random 90-degree rotations, and image mirroring during training to prevent the GCN from over-fitting.

Since we employed two different graph convolution paradigms, namely spatial-based and spectral-based, the results are categorized accordingly. It is worth noting that we used PyTorch Geometric library to implement the graph-based models [72].

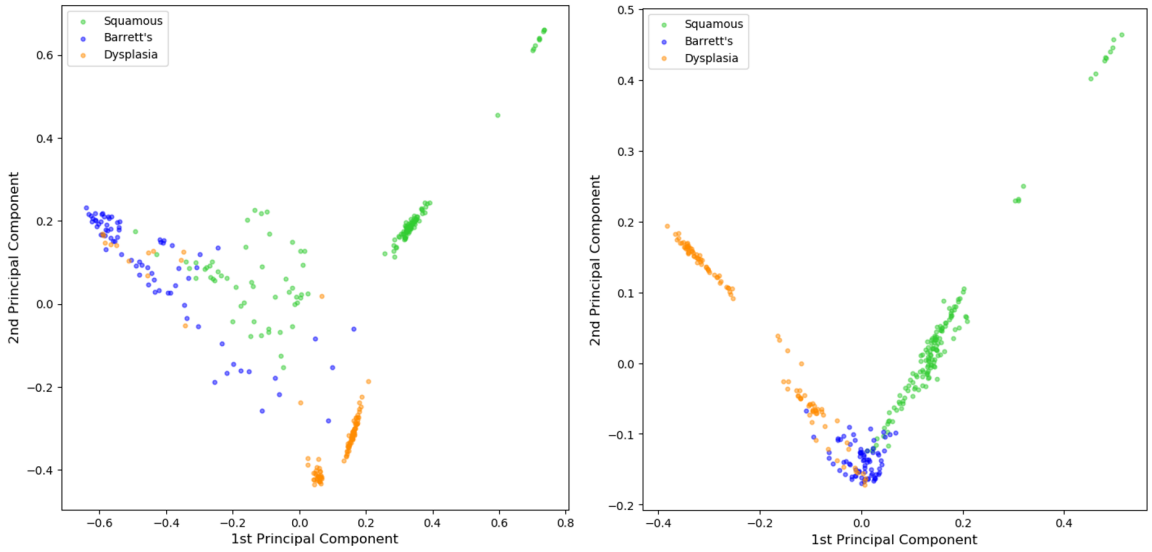

Figure 6.3: PCA plot for WSIs encoded using (left) SGConv-CAE, (right) SGConv-BiGAN.

Spatial-based GCNs

For spatial-based convolution layers, we used SGConv [73]. Figure 6.3 shows the results of PCA that visualize the image-level representations. As illustrated, WSIs in squamous and dysplastic BE classes have been encoded relatively separately from each other, while non-dysplastic BE WSIs have sort of confusion with both of these classes. This pattern was also observed in previous models in this study.

Classification results can further refine our findings from the PCA plots. To evaluate the model quantitatively, five standard metrics were used for classification under a 1-vs-rest strategy: accuracy, precision, recall, specificity, and F1 score. Bootstrapping was used for all metrics to estimate the results' significance.

The results of employing a spatial-based GCN model for the classification of esophageal WSIs in three classes of squamous, dysplastic BE, and non-dysplastic BE are summarized in Table 6.1. The reported values are averages with 95% confidence intervals.

As shown in Table 6.1, the model on CAE-derived image features outperforms the model on BiGAN-based ones given the weighted average F1 score. The weighted average of F1 score for the model trained on encoded WSIs using CAE is 0.876 (95% CI, 0.873-0.880) vs. 0.853 (95% CI, 0.838-0.867) for the model trained on encoded WSIs using BiGAN. Since two 95% confidence intervals do not overlap, the null hypothesis of zero difference between their F1 scores is rejected at the 0.05 level. Tables 6.2 and 6.3 shows the confusion matrix of SGConv-CAE and SGConv-BiGAN respectively. As shown in both models, dysplastic BE has mostly confusion with non-dysplastic BE.

Table 6.1: Results of WSI classification using CBGCN (SGConv) approach

Class	Metric	Model	
		SGConv-CAE	**SGConv-BiGAN**
Squamous	Accuracy	0.922 (0.919, 0.926)	0.908 (0.894, 0.922)
	Precision	0.993 (0.991, 0.995)	0.960 (0.951, 0.968)
	Recall	0.831 (0.824, 0.838)	0.829 (0.798, 0.860)
	Specificity	0.995 (0.994, 0.996)	0.971 (0.965, 0.977)
	F1 score	0.904 (0.900, 0.908)	0.881 (0.861, 0.900)
Barrett's	Accuracy	0.874 (0.870, 0.878)	0.848 (0.832, 0.863)
	Precision	0.649 (0.639, 0.659)	0.657 (0.629, 0.686)
	Recall	0.967 (0.963, 0.972)	0.855 (0.839, 0.871)
	Specificity	0.846 (0.841, 0.851)	0.845 (0.824, 0.866)
	F1 score	0.776 (0.768, 0.783)	0.732 (0.712, 0.752)
Dysplasia	Accuracy	0.942 (0.939, 0.944)	0.937 (0.933, 0.940)
	Precision	0.968 (0.964, 0.973)	0.939 (0.934, 0.944)
	Recall	0.853 (0.845, 0.860)	0.864 (0.853, 0.875)
	Specificity	0.986 (0.984, 0.988)	0.972 (0.970, 0.975)
	F1 score	0.906 (0.902, 0.910)	0.899 (0.892, 0.905)
Weighted Average	Accuracy	0.917 (0.915, 0.920)	0.903 (0.893, 0.914)
	Precision	0.907 (0.905, 0.910)	0.884 (0.876, 0.892)
	Recall	0.869 (0.865, 0.873)	0.846 (0.831, 0.862)
	Specificity	0.958 (0.957, 0.960)	0.942 (0.937, 0.948)
	F1 score	0.876 (0.873, 0.880)	0.853 (0.838, 0.867)

Table 6.2: Confusion matrix of SGConv-CAE

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	124 (0.873)	18 (0.127)	0 (0.000)
	Non-dysplastic BE	1 (0.014)	69 (0.932)	4 (0.054)
	Dysplastic BE	7 (0.067)	12 (0.114)	86 (0.819)

Figure 6.4 summarizes the results of the evaluation of different numbers of clusters for both CAE and BiGAN approaches. As shown, the SGConv-CAE outperforms the SGconv-BiGAN regardless of number of clusters. Furthermore, by changing the number of clusters, there is no significant change in the performance of the SGConv-CAE. In contrast, the number of clusters is a more critical parameter in SGConv-BiGAN. For both SGConv-CAE and SGConv-BiGAN, we set the number of clusters to 100 and the results summarized in Tables 6.1, 6.2 and 6.3 are based on 100 clusters for both models.

Table 6.3: Confusion matrix of SGConv-BiGAN

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	118 (0.831)	23 (0.162)	1 (0.007)
	Non-dysplastic BE	1 (0.014)	62 (0.932)	4 (0.054)
	Dysplastic BE	2 (0.019)	22 (0.210)	81 (0.771)

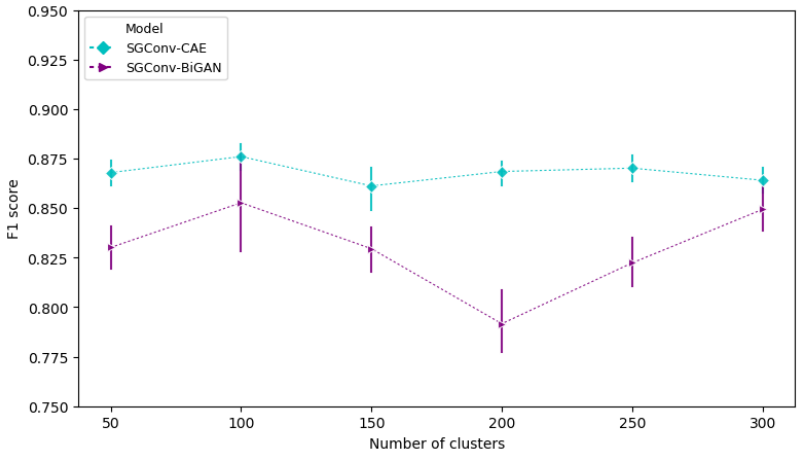

Figure 6.4: The effect of the number of clusters on the performance of CBGCN (SGConv)

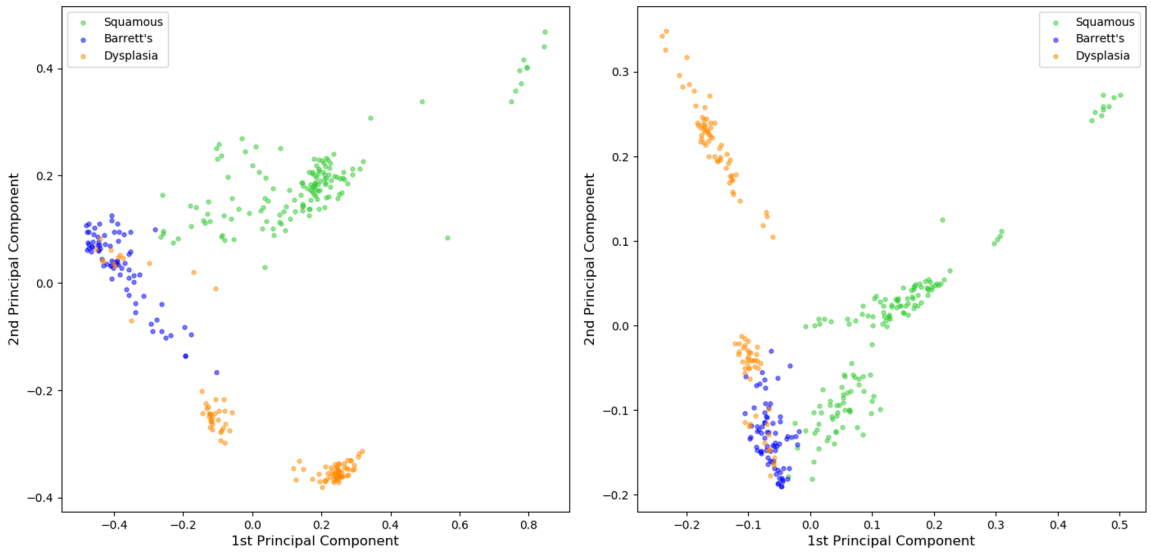

Figure 6.5: PCA plot for WSIs encoded using (left) SGConv-CAE, (right) SGConv-BiGAN

Spectral-based GCNs

For spectral-based convolution layers, we used ChebConv [70]. Figure 6.5 shows the results of PCA on WSIs encoded by ChebConv-based CBGCN. Same as previous models, including the spatial-based approach, the WSIs in squamous and dysplastic BE have been encoded relatively separately. In contrast, non-dysplastic BE WSIs have confusion with both of these classes.

The classification results of esophageal WSIs for different models are summarized in Table 6.4. The reported values are averages with 95% confidence intervals. As shown, the weighted average of F1 score for the model trained on encoded WSIs using CAE is 0.883 (95%

Table 6.4: Results of WSI classification using CBGCN (ChebConv) approach

Class	Metric	Model	
		ChebConv-CAE	**ChebConv-BiGAN**
Squamous	Accuracy	0.928 (0.923, 0.934)	0.924 (0.921, 0.927)
	Precision	0.981 (0.976, 0.986)	0.999 (0.999, 1.000)
	Recall	0.854 (0.843, 0.864)	0.828 (0.823, 0.834)
	Specificity	0.987 (0.984, 0.991)	0.999 (0.999, 1.000)
	F1 score	0.912 (0.905, 0.920)	0.906 (0.902, 0.909)
Barrett's	Accuracy	0.885 (0.880, 0.890)	0.871 (0.868, 0.875)
	Precision	0.675 (0.663, 0.687)	0.646 (0.638, 0.654)
	Recall	0.971 (0.965, 0.977)	0.976 (0.973, 0.979)
	Specificity	0.859 (0.852, 0.867)	0.839 (0.835, 0.844)
	F1 score	0.794 (0.786, 0.802)	0.777 (0.770, 0.783)
Dysplasia	Accuracy	0.941 (0.939, 0.944)	0.947 (0.945, 0.950)
	Precision	0.976 (0.971, 0.980)	0.982 (0.979, 0.984)
	Recall	0.844 (0.836, 0.852)	0.854 (0.847, 0.861)
	Specificity	0.989 (0.987, 0.991)	0.992 (0.991, 0.993)
	F1 score	0.904 (0.900, 0.909)	0.913 (0.909, 0.917)
Weighted Average	Accuracy	0.923 (0.919, 0.927)	0.919 (0.917, 0.921)
	Precision	0.910 (0.906, 0.914)	0.913 (0.911, 0.914)
	Recall	0.877 (0.871, 0.883)	0.871 (0.868, 0.875)
	Specificity	0.959 (0.956, 0.961)	0.960 (0.959, 0.962)
	F1 score	0.883 (0.878, 0.889)	0.879 (0.876, 0.882)

Table 6.5: Confusion matrix of ChebConv-CAE

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	133 (0.937)	9 (0.063)	0 (0.000)
	Non-dysplastic BE	0 (0.000)	73 (0.986)	1 (0.014)
	Dysplastic BE	13 (0.124)	13 (0.124)	79 (0.752)

CI, 0.878-0.889) vs. 0.879 (95% CI, 0.876-0.882) for the model trained on encoded WSIs using BiGAN. The p-value of comparing two models is 0.029, which means the ChebConv model on CAE-derived features significantly outperforms the ChebConv on BiGAN-based features. Tables 6.5 and 6.6 shows the confusion matrix of ChebConv-CAE and ChebConv-BiGAN respectively.

Figure 6.6 summarizes the results of the evaluation of different numbers of clusters for both the CAE and BiGAN approaches. As shown, the ChebConv-CAE performs better than the Chebconv-BiGAN regardless of the number of clusters. Furthermore, unlike

Table 6.6: Confusion matrix of ChebConv-BiGAN

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	114 (0.803)	28 (0.197)	0 (0.000)
	Non-dysplastic BE	1 (0.014)	69 (0.932)	4 (0.054)
	Dysplastic BE	0 (0.000)	14 (0.133)	91 (0.867)

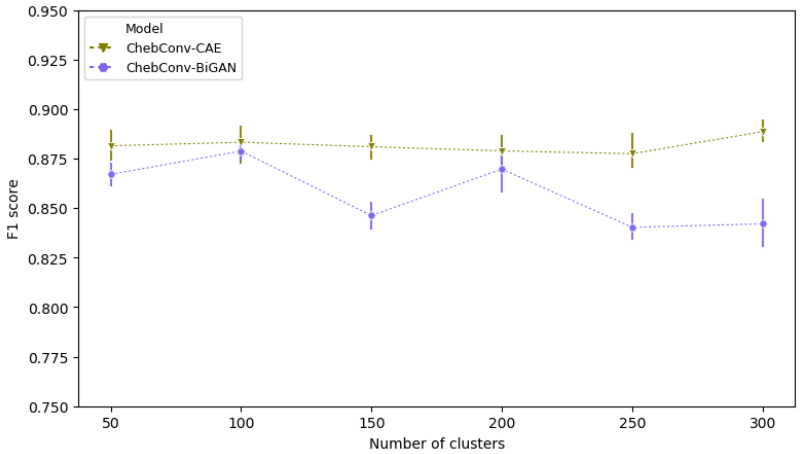

Figure 6.6: The effect of the number of clusters on the performance of CBGCN (ChebConv)

the ChebConv-BiGAN, the model on CAE-derived features is more stable to changes in the number of clusters. For both ChebConv-CAE and ChebConv-BiGAN, 100 clusters is determined as an optimal value for number of clusters and the results summarized in Tables 6.4, 6.5 and 6.6 are accordingly.

Spatial-based vs spectral-based CBGCN

Generally speaking, the classification performance of spectral-based GCN on our dataset is better than the performance of spatial-based GCN (see Figure 6.7). However, the difference in performance varies depending on the number of clusters. Since the best performance of these models is on number of clusters equal to 100, we limit our comparison to this value. The F1 scores from Tables 6.1 and 6.4 are summarized in Table 6.7.

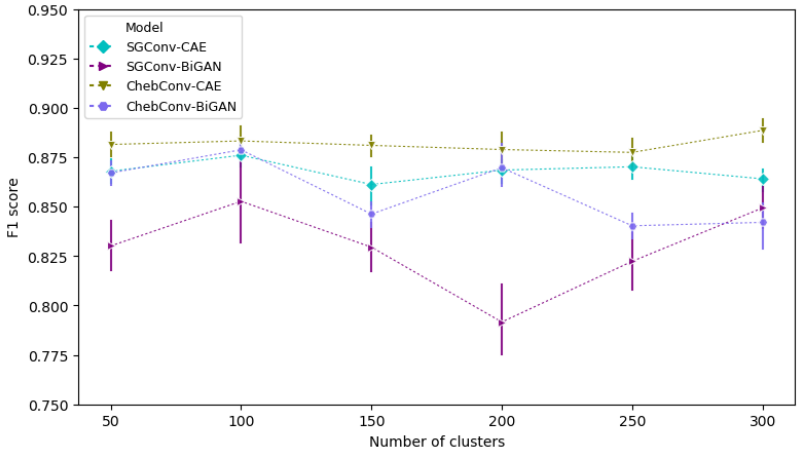

Figure 6.7: Comparison of the performance of spatial-based and spectral-based CBGCN

Table 6.7: Comparison of the performance of spatial-based and spectral-based CBGCN

	CAE-derived features	BiGAN-derived features
Spatial-based (SGConv)	0.876 (0.873, 0.880)	0.853 (0.838, 0.867)
Spectral-based (ChebConv)	0.883 (0.878, 0.889)	0.879 (0.876, 0.882)

As indicated, the spectral-based model outperforms the spatial-based one given both feature extracting approaches. However on BiGAN-derived features this superiority is significant at 0.05 level ($p-value \approx 0.000$) but on CAE-based feature, it is not ($p-value \approx 0.033$).

CBGCN vs GCN

The contribution of CBGCN over GCN is incorporating the concept of visual words into the model by encoding the WSIs using visual words rather than applying patch-level embedding vectors directly. Tables 6.8 and 6.9 show the confusion matrices of GCN (ChebConv) given the CAE-derived and BiGAN-derived image features respectively.

Table 6.8: Confusion matrix of ChebConv-CAE (without clustering)

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	101 (0.711)	34 (0.239)	7 (0.049)
	Non-dysplastic BE	3 (0.041)	71 (0.995)	0 (0.000)
	Dysplastic BE	5 (0.048)	41 (0.390)	59 (0.562)

Table 6.9: Confusion matrix of ChebConv-BiGAN (without clustering)

		Predicted label		
		Squamous	**Non-dysplastic BE**	**Dysplastic BE**
True label	**Squamous**	113 (0.796)	29 (0.204)	0 (0.000)
	Non-dysplastic BE	1 (0.014)	70 (0.946)	3 (0.041)
	Dysplastic BE	0 (0.000)	44 (0.419)	61 (0.581)

Comparing these results with the results of the CBGCN model (Tables 6.5 and 6.6), it can be seen that the cluster-based GCN outperforms the GCN model for the classification of esophageal WSIs. We employed bootstrapping approach (100 iterations) to test the significance of the results. Figure 6.8, represents the boxplots of F1 score values on different iterations for both GCN and CBGCN models.

As shown, the performance of CBGCN is significantly better than GCN. As a result of employing the cluster-based model, the most improvement has been in the classification of dysplastic BE WSIs. This class has a great confusion with non-dysplastic BE when we

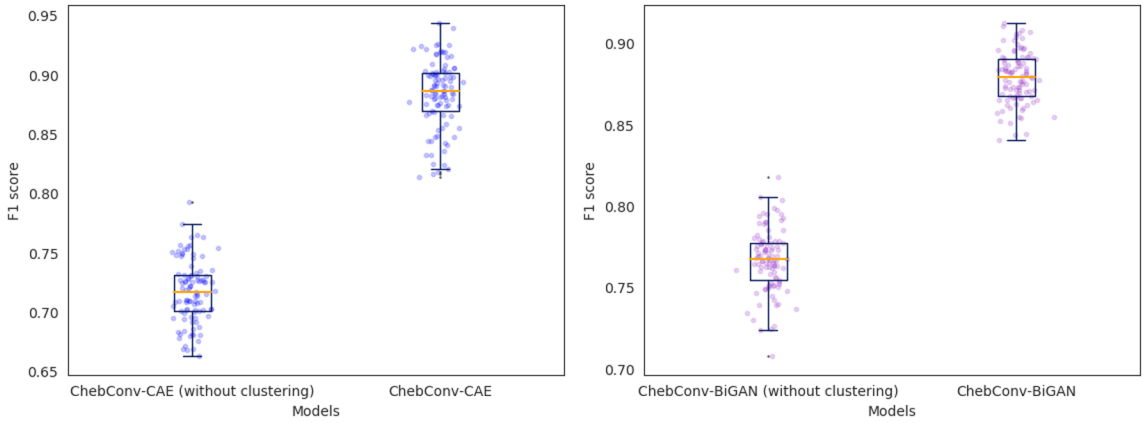

Figure 6.8: Comparison of the performance of CBGCN (ChebConv) and GCN (ChebConv) for classification of WSIs encoded by (left) CAE, (right) BiGAN

Table 6.10: Comparison of the performance of BoVW, CBNIC, and CBGCN

	CAE-derived features	**BiGAN-derived features**
BoVW	0.812 (0.808, 0.816)	0.811 (0.807, 0.815)
CBNIC	0.889 (0.884, 0.894)	0.858 (0.855, 0.862)
CBGCN (Spatial-based)	0.876 (0.873, 0.880)	0.853 (0.838, 0.867)
CBGCN (Spectral-based)	0.883 (0.878, 0.889)	0.879 (0.876, 0.882)

use GCN to classify them.

CBGCN vs CBNIC and BoVW

In this study, the CBNIC and CBGCN have been proposed to provide an image-level representation taking advantage of visual words concept and also preserving the spatial arrangement of local representations, a characteristic that the BoVW lacks. The same visual words were employed for image encoding in all models to make sure that the clustering effect is the same on all models. Figure 6.9 illustrates the performance of all models investigated in this study given the different number of clusters. For more detailed analysis, the best performance of different models (considering both accuracy and interpretability) has been summarized in Table 6.10. As shown, both CBNIC and CBGCN outperform the BoVW model. Furthermore, the models on CAE-derived image features perform better compared to BiGAN-derived features.

Analysis of Codeblocks

After training the classifier on compressed images, a weight is learned for each codeblock that is an estimate of how important it is in the image-level inference. By aggregating the weights from all images in the test set for every codeblock and averaging, a value is obtained

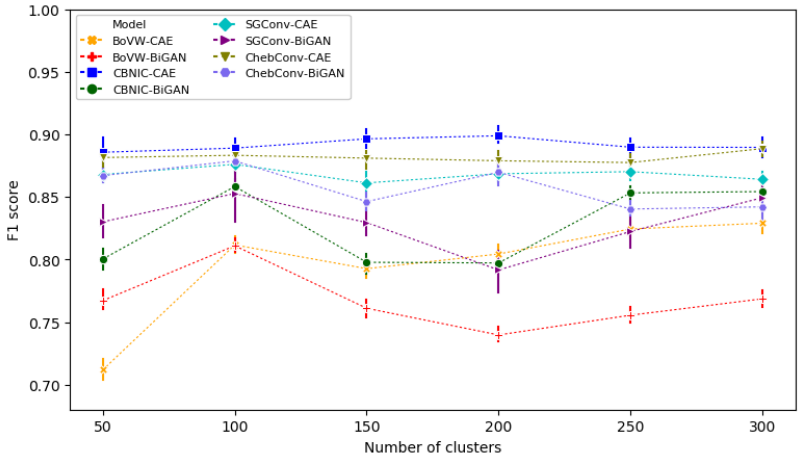

Figure 6.9: Comparison of the performance of CBGCN, CBNIC and BoVW

indicating the relative importance of the given codeblock. Figures 6.10 and 6.11 show some randomly selected image patches associated with codeblocks of the highest importance for CAE-derived and BiGAN-derived features, respectively.

A medical expert reviewed these sampled tissue tiles. In both models, squamous tissue tiles were indicative of normal esophageal histology showing squamous cells that are flat, thin cells lining the esophagus's surface. For non-dysplastic BE, tiles showed goblet cells and columnar epithelium typically only present in the small intestine, but with Barrett's, the normal esophageal lining shows the presence of these mucin-containing goblet cells and columnar epithelial cells. For dysplastic BE, crowded hyperchromatic nuclei were visualized that are known to be present among pre-cancerous dysplastic cells.

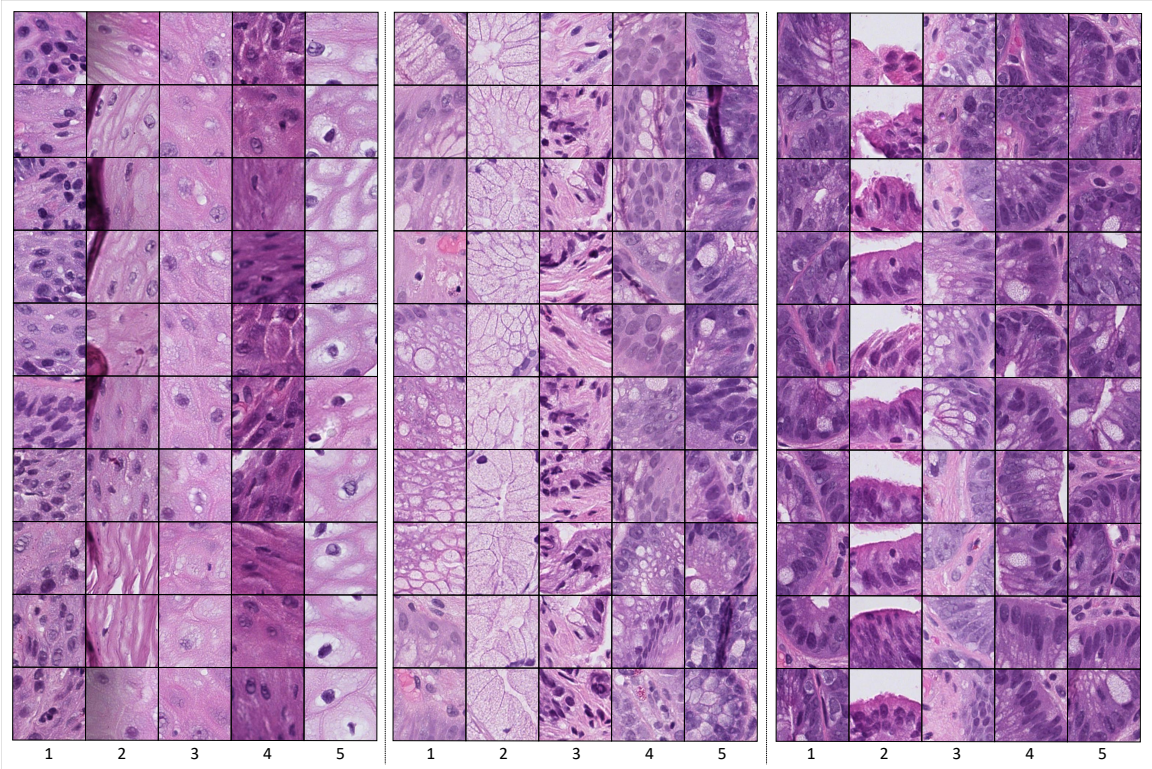

Figure 6.10: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the CBGCN-CAE

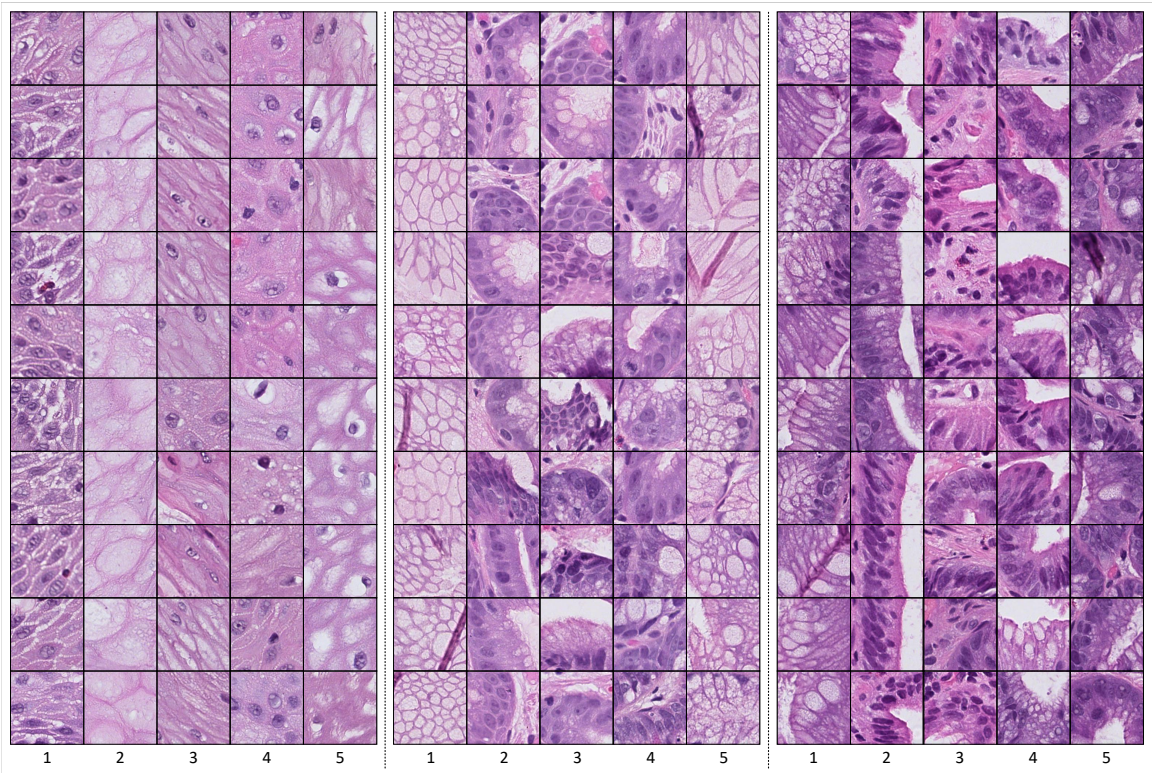

Figure 6.11: Randomly sampled tissue tiles from top 5 codeblocks associated with (left) Squamous, (middle) Barrett's, and (right) Dysplasia in the CBGCN-BiGAN

6.5 Discussion

In this chapter, we employed a graph convolution neural network to classify esophageal WSIs. This approach learns an image-level representation preserving the spatial arrangement of visual words associated with image patches. The experimental results confirmed the ability of this model to diagnose non-dysplastic and dysplastic BE. As demonstrated, the cluster-based GCN outperforms the GCN trained on compressed WSIs considering the tile-level representations, which indicates the positive contribution of encoding the WSIs using visual words. Employing visual words rather than tile-level representations also enhances the model interpretability. In the proposed model, we employed an attention pooling layer to estimate each visual word's importance in the predicted label for each image. By aggregating these values from the individual images, we calculated the overall importance of each visual word. From a medical perspective, pixel-wise region localization can provide an accurate visual explanatory factor for the model's performance, a highly desirable property in decision support systems. The qualitative performance of this model can be evaluated simply by inspecting the top-rank visual words by a pathologist. Generally speaking, forasmuch as the CBCNG deals with a visual dictionary containing a finite number of words, the results are far more interpretable than the model dealing with a massive number of tile-level representations.

Although in this chapter, we employed CBGCN on the same size images, the graph-based encoding of gigapixel images can be employed on images of different sizes. This is an advantage of graph neural networks over the NIC approach for encoding gigapixel images. Furthermore, the inherent flexibility of graph-based structures makes it possible to examine diverse spatial patterns.

7 | Conclusions and Future Work

Whole-slide histopathology images play a crucial role in tissue specimen assessment and exploring underlying mechanisms associated with disease progression and patient outcome. Visual inspection of these images by pathologists is labor-intensive, time-consuming, and subject to high inter-observer variability. It has created a need for automated extraction of meaningful information from WSIs using artificial intelligence and machine learning techniques. In this dissertation, we investigated the ability of three different approaches (i.e., BoVW, CBNIC, and CBGCN) to provide a representation of whole-slide histopathology images. The quality of image representations was evaluated on an esophageal dataset. The proposed approaches were employed to classify and locate non-dysplastic BE and dysplastic BE on histopathology images and achieved promising results.

The BoVW approach encodes the images as an order-less histogram of visual word frequencies. This model provides an image-wise representation in a totally unsupervised way. This model utilizes neither annotated images nor image labels. Although BoVW showed a satisfactory performance on the classification of esophageal WSIs, it ignores the spatial arrangement of visual words, which leads to some information loss. In some cases, such information loss leads to significant deterioration of model performance.

The CBNIC and CBGCN were proposed to provide WSI-level representation preserving spatial arrangement of local image features and leveraging visual words that have already shown a promising contribution in the WSI encoding in the BoVW. The CBNIC provides a compressed representation of WSIs by organizing the visual words associated with local tissue tiles derived from a uniform grid of square patches. This approach sidesteps the steep computational requirements needed to train neural networks on whole slide images by creating highly compact representations of gigapixel images that are more amenable to training neural networks.

The CBGCN utilizes graph neural networks for image representation. This model encodes the WSIs as a graph in which the nodes are visual words connected using vertexes to capture their spatial proximity information. Training a graph convolutional network on WSI-level graphs outputs the label of new images.

The results of the classification of esophageal WSIs showed the superiority of CBNIC and CBGCN over BoVW. Although codeblock analysis was accomplished for all models, no comparative analysis was performed between the top image patterns of different models. In future research, these results should be investigated from a medical perspective to see whether employing CBNIC and CBGCN has led to more informative patterns than BoVW.

In all models studied in this dissertation, the tile-level representations extracted by CAE or BiGAN were clustered to generate visual words in an unsupervised fashion. The results demonstrated the capability of unsupervised approach for extracting relevant image features from WSIs, and consequently, better identifying dysplastic and non-dysplastic BE. This is important because these approaches avoid the need for manual examination of images.

A deep learning-based model for detecting and locating dysplastic and non-dysplastic BE patterns on histopathologic images has a wide variety of applications in clinical settings. Such a model can be integrated into clinical information management systems as a decision support system. Such systems can provide clinicians and practitioners with possible diagnoses or improve confidence in their assessments via providing second opinions for prognostic decision-making of more challenging histopathological patterns. Successful implementation of this system can support a more accurate classification of pre-malignant diseases of the esophagus.

Gastroenterologists obtain some esophageal tissue samples from each patient and examine them to diagnose BE severity. Nevertheless, the models proposed in this dissertation predict WSI-level labels. This gap can be addressed in future research by extending these models to aggregate image-level predictions from all samples of a patient and predict the patient outcome.

Albeit achieving promising results, this study has some general limitations as well as model-specific ones. First, all biopsy images used for this study were collected from a single center and scanned with the same equipment. Thus, such data might not be representative of the entire range of histological patterns in patients worldwide. Collaboration with other medical centers and collecting more images would refine our model using a more diverse dataset. Second, to have enough images of the same size to train the model (since in the CBNIC model, a CNN is trained on compressed WSIs, same size images is a requirement for this model, while BoVW and graph-based models can be trained on images of different size), we had to generate images with size 5000×5000. To better evaluate the models' capability, a larger dataset and larger images would be ideally preferred. Third, in the proposed models, we divided WSIs into small patches covering 200×200 pixels to be fed into a feature extractor. The size of patches might be a critical parameter and drastically impacts the model performance. The performance of the models with varying sizes of tissue tiles would be worthwhile to investigate. Fourth, dealing with imbalanced datasets (the dataset we

employed in this study is not the case), the visual codebook is likely to be dominated by the most observed types of texture samples. Employing multiple disjoint dictionaries [37] can avoid this issue. Fifth, in this study, we applied a single method of stain normalization, and the use of other methods may lead to different results. Therefore, investigating the effect of different stain normalization techniques can be another potential area of future work. The last but not the least, the experimental results in this study are based on feature extraction from tissue tiles. A patch-based encoding scheme cannot provide experts and clinicians with a cell-level insight that is critical in some diseases. Encoding the histology images given the cell morphology and cell organization as a graph to capture the tissue information [41, 64] and employing a graph-based model on cell-level graphs can address this issue. This also can be considered as an avenue for future research.

Some findings of this dissertation have been published in a couple of papers. Some more papers are also being written that apply the methods presented in this study to diagnose some other gastrointestinal diseases. They have been listed in the Table 7.1.

Table 7.1: List of papers

No.	Paper	Staus
1	**R. Sali**, N. Moradinasab, S. Gluria, L. Ehsan, P. Fernandes, T. U. Shah, S. Syed and D. Brown, "Deep Learning for Whole-Slide Tissue Histopathology Classification: A Comparative Study in the Identification of Dysplastic and Non-Dysplastic Barrett's Esophagus," Journal of Personalized Medicine, vol. 10, no. 4, 2020.	Published
2	S. Guleria, T. U. Shah, J. V. Pulido, M. Fasullo, L. Ehsan, R. Lippman, **R. Sali**, P. Mutha, L. Cheng, D. E. Brown and S. Syed, "Deep learning systems detect dysplasia with human-like accuracy using histopathology and probe-based confocal laser endomicroscopy," Scientific reports, vol. 11, no. 1, pp. 1-11, 2021.	Published
3	**R. Sali**, L. Ehsan, S. Guleria, T. U. Shah, M. Fasullo, R. Lippman, P. Mutha, S. Syed and D. E. Brown, "CBNIC: Cluster-Based Neural Image Compression for Representation Learning of Whole Slide Histopathology Images,"	Under Review
4	Prediction of Celiac Disease Severity and Associated Endocrine Morbidities on Whole-Slide Histopathology Images through Deep Learning-based Image Analytics	Working paper
5	Deep Graph Neural Networks for Crohn's Disease Diagnosis on Whole-Slide Tissue Histopathology Images	Working paper

The subject of one of our working papers is the prediction of Celiac Disease (CD) severity on whole-slide histopathology images using deep learning approaches. This study is an extension of CeliacNet [10], one of our previously published works in the Gastroenterology Data Science Lab at the University of Virginia. CeliacNet employs a deep learning model to predict CD's severity (based on Marsh score) on tissue tiles using a weakly-supervised approach. In this model, the results of patch classification are aggregated for making an inference about the WSIs. In the second working paper, we train the deep models proposed in this dissertation on ileal biopsies from subjects with distinct Crohn's phenotype and histologically controls. This study aims to predict different classes of Crohn's disease,

including inflammatory (B1) children at diagnosis who maintained B1 behavior or went on to develop stricturing (B2), penetrating (B3), or both (B2/B3) subtypes on ileal WSIs.

References

[1] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[2] S. J. Spechler, "Barrett's esophagus," *New England Journal of Medicine*, vol. 346, no. 11, pp. 836–842, 2002.

[3] N. J. Shaheen, G. W. Falk, P. G. Iyer, and L. B. Gerson, "Acg clinical guideline: diagnosis and management of barrett's esophagus," *American Journal of Gastroenterology*, vol. 111, no. 1, pp. 30–50, 2016.

[4] G. W. Falk, T. W. Rice, J. R. Goldblum, and J. E. Richter, "Jumbo biopsy forceps protocol still misses unsuspected cancer in barrett's esophagus with high-grade dysplasia," *Gastrointestinal endoscopy*, vol. 49, no. 2, pp. 170–176, 1999.

[5] R. Kariv, T. P. Plesec, J. R. Goldblum, M. Bronner, M. Oldenburgh, T. W. Rice, and G. W. Falk, "The seattle protocol does not more reliably predict the detection of cancer at the time of esophagectomy than a less intensive surveillance protocol," *Clinical Gastroenterology and Hepatology*, vol. 7, no. 6, pp. 653–658, 2009.

[6] S. Wani, J. H. Rubenstein, M. Vieth, and J. Bergman, "Diagnosis and management of low-grade dysplasia in barrett's esophagus: expert review from the clinical practice updates committee of the american gastroenterological association," *Gastroenterology*, vol. 151, no. 5, pp. 822–835, 2016.

[7] P. Vennalaganti, V. Kanakadandi, J. R. Goldblum, S. C. Mathur, D. T. Patil, G. J. Offerhaus, S. L. Meijer, M. Vieth, R. D. Odze, S. Shreyas, *et al.*, "Discordance among pathologists in the united states and europe in diagnosis of low-grade dysplasia for patients with barrett's esophagus," *Gastroenterology*, vol. 152, no. 3, pp. 564–570, 2017.

[8] E. Downs-Kelly, J. E. Mendelin, A. E. Bennett, E. Castilla, W. H. Henricks, L. Schoenfield, M. Skacel, L. Yerian, T. W. Rice, L. A. Rybicki, *et al.*, "Poor interobserver agreement in the distinction of high-grade dysplasia and adenocarcinoma in pretreatment barrett's esophagus biopsies," *American Journal of Gastroenterology*, vol. 103, no. 9, pp. 2333–2340, 2008.

[9] J. Rony, S. Belharbi, J. Dolz, I. B. Ayed, L. McCaffrey, and E. Granger, "Deep weakly-supervised learning methods for classification and localization in histology images: a survey," *arXiv preprint arXiv:1909.03354*, 2019.

[10] R. Sali, L. Ehsan, K. Kowsari, M. Khan, C. A. Moskaluk, S. Syed, and D. E. Brown, "Celiacnet: Celiac disease severity diagnosis on duodenal histopathological images using deep residual networks," *arXiv preprint arXiv:1910.03084*, 2019.

[11] R. Sali, S. Adewole, L. Ehsan, L. A. Denson, P. Kelly, B. C. Amadi, L. Holtz, S. A. Ali, S. R. Moore, S. Syed, *et al.*, "Hierarchical deep convolutional neural networks for multi-category diagnosis of gastrointestinal disorders on histopathological images," *arXiv preprint arXiv:2005.03868*, 2020.

[12] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2424–2433, 2016.

[13] 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.

[14] 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.

[15] D. Tellez, J. van der Laak, and F. Ciompi, "Gigapixel whole-slide image classification using unsupervised image compression and contrastive training," 2018.

[16] Y. Huang and A. C.-s. Chung, "Improving high resolution histology image classification with deep spatial fusion network," in *Computational Pathology and Ophthalmic Medical Image Analysis*, pp. 19–26, Springer, 2018.

[17] S. Shafiei, A. Safarpoor, A. Jamalizadeh, and H. Tizhoosh, "Class-agnostic weighted normalization of staining in histopathology images using a spatially constrained mixture model," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3355–3366, 2020.

[18] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[19] R. Sparks and A. Madabhushi, "Explicit shape descriptors: Novel morphologic features for histopathology classification," *Medical image analysis*, vol. 17, no. 8, pp. 997–1009, 2013.

[20] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *International journal of computer vision*, vol. 113, no. 1, pp. 3–18, 2015.

[21] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 228–235, Springer, 2014.

[22] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE transactions on medical imaging*, vol. 35, no. 3, pp. 738–751, 2015.

[23] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*, vol. 9041, p. 904103, International Society for Optics and Photonics, 2014.

[24] J. W. Wei, L. J. Tafe, Y. A. Linnik, L. J. Vaickus, N. Tomita, and S. Hassanpour, "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[25] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep learning for classification of colorectal polyps on whole-slide images," *Journal of pathology informatics*, vol. 8, 2017.

[26] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, A. Madabhushi, and F. González, "High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: Application to invasive breast cancer detection," *PloS one*, vol. 13, no. 5, p. e0196828, 2018.

[27] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised learning for whole slide lung cancer image classification," 2018.

[28] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019.

[29] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[30] Y. Yamamoto, T. Tsuzuki, J. Akatsuka, M. Ueki, H. Morikawa, Y. Numata, T. Takahara, T. Tsuyuki, K. Tsutsumi, R. Nakazawa, *et al.*, "Automated acquisition of explainable knowledge from unannotated histopathology images," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[31] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 126–135, Springer, 2009.

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

[33] P. Alirezazadeh, B. Hejrati, A. Monsef-Esfahani, and A. Fathi, "Representation learning-based unsupervised domain adaptation for classification of breast cancer histopathology images," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 3, pp. 671–683, 2018.

[34] V. Popovici, E. Budinská, L. Čápková, D. Schwarz, L. Dušek, J. Feit, and R. Jaggi, "Joint analysis of histopathology image features and gene expression in breast cancer," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–9, 2016.

[35] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1626–1630, IEEE, 2014.

[36] R. Sali, N. Moradinasab, S. Guleria, L. Ehsan, P. Fernandes, T. U. Shah, S. Syed, and D. E. Brown, "Deep learning for whole-slide tissue histopathology classification: A comparative study in the identification of dysplastic and non-dysplastic barrett's esophagus," *Journal of Personalized Medicine*, vol. 10, no. 4, p. 141, 2020.

[37] S. Zhu, Y. Li, S. Kalra, and H. R. Tizhoosh, "Multiple disjoint dictionaries for representation of histopathology images," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 243–252, 2018.

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

[39] A. Foncubierta-Rodríguez, A. García Seco de Herrera, and H. Müller, "Medical image retrieval using bag of meaningful visual words: unsupervised visual vocabulary pruning with plsa," in *Proceedings of*

the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, pp. 75–82, 2013.

[40] M. Adnan, S. Kalra, and H. R. Tizhoosh, "Representation learning of histopathology images using graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 988–989, 2020.

[41] P. Pati, G. Jaume, A. Foncubierta, F. Feroce, A. M. Anniciello, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, *et al.*, "Hierarchical cell-to-tissue graph representations for breast cancer subtyping in digital pathology," *arXiv preprint arXiv:2102.11057*, 2021.

[42] S. Gadiya, D. Anand, and A. Sethi, "Histographs: Graphs in histopathology," *arXiv preprint arXiv:1908.05020*, 2019.

[43] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, 2020.

[44] H. Sharma, N. Zerbe, S. Lohmann, K. Kayser, O. Hellwich, and P. Hufnagl, "A review of graph-based methods for image analysis in digital histopathology," *Diagnostic pathology*, vol. 1, no. 1, 2015.

[45] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[46] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides," *JAMA network open*, vol. 2, no. 11, pp. e1914645–e1914645, 2019.

[47] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*, pp. 2127–2136, PMLR, 2018.

[48] T. Shah, R. Lippman, D. Kohli, P. Mutha, S. Solomon, and A. Zfass, "Accuracy of probe-based confocal laser endomicroscopy (pcle) compared to random biopsies during endoscopic surveillance of barrett's esophagus," *Endoscopy international open*, vol. 6, no. 04, pp. E414–E420, 2018.

[49] https://nanozoomer.hamamatsu.com/jp/en/product/search/C13220-01/index.html.

[50] K. Kowsari, R. Sali, M. N. Khan, W. Adorno, S. A. Ali, S. R. Moore, B. C. Amadi, P. Kelly, S. Syed, and D. E. Brown, "Diagnosis of celiac disease and environmental enteropathy on biopsy images using color balancing on convolutional neural networks," in *Proceedings of the Future Technologies Conference*, pp. 750–765, Springer, 2019.

[51] F. G. Zanjani, S. Zinger, *et al.*, "Deep convolutional gaussian mixture model for stain-color normalization of histopathological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 274–282, Springer, 2018.

[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[53] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[54] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words," *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 733–746, 2010.

[55] T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. B. Wein, "Content-based image retrieval in medical applications," *Methods of information in medicine*, vol. 43, no. 04, pp. 354–361, 2004.

[56] R. T. Powell, A. Olar, S. Narang, G. Rao, E. Sulman, G. N. Fuller, and A. Rao, "Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: a preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of the cancer genome atlas," *Journal of pathology informatics*, vol. 8, 2017.

[57] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *Ieee Access*, vol. 6, pp. 24680–24693, 2018.

[58] K. Zhang, H. Zhou, L. Chen, M. Fei, J. Wu, and P. Zhang, "A novel segmentation framework using sparse random feature in histology images of colon cancer," in *Advanced Computational Methods in Life System Modeling and Simulation*, pp. 173–180, Springer, 2017.

[59] H. Mittal and M. Saraswat, "Classification of histopathological images through bag-of-visual-words and gravitational search algorithm," in *Soft Computing for Problem Solving*, pp. 231–241, Springer, 2019.

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

[61] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[63] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[64] M. Sureka, A. Patil, D. Anand, and A. Sethi, "Visualization for histopathology images using graph convolutional neural networks," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 331–335, IEEE, 2020.

[65] Y. Zhou, S. Graham, N. Alemi Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[66] R. Konda, H. Wu, and M. D. Wang, "Graph convolutional neural networks to classify whole slide images," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 754–758, IEEE, 2020.

[67] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, pp. 420–434, Springer, 2001.

[68] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*, pp. 1263–1272, PMLR, 2017.

[69] M. Balcilar, G. Renton, P. Héroux, B. Gauzere, S. Adam, and P. Honeine, "Bridging the gap between spectral and spatial domains in graph neural networks," *arXiv preprint arXiv:2003.11702*, 2020.

[70] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *arXiv preprint arXiv:1606.09375*, 2016.

[71] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[72] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.

[73] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International conference on machine learning*, pp. 6861–6871, PMLR, 2019.